SILENCE AND VOICED-UNVOICED-MIXED EXCITATION CLASSIFICATION OF SPEECH WITH APPLICATIONS: A TWO-CHANNEL AND A ONE-CHANNEL

By

MINSOO HAHN

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPY

UNIVERSITY OF FLORIDA

1989

To my parents,

to my wife and son,

Special Thanks.

ACKNOWLEDGEMENTS

The completion of this research has confirmed that the gratitude traditionally extended on this page is indeed sincere. First and foremost, thanks are due to the author's advisor and committee chairman, Dr. D.G. Childers, for his encouragement and invaluable assistance throughout this research. The author also thanks Drs. J.T. Tou, N.W. Perry, Jr., A.A. Arroyo, and J.C. Principe for their time and interest in serving on the supervisory committee. Thanks are also due to Dr. H.B. Rothman for his help throughout this work.

The author wishs to thank all my colleagues at the Mind-Machine Interaction Research Center for their help and advice. He also thanks Mr. T.R. Sawallis for being a cooperative corrector and reviewer.

Finally, he wishes to express appreciation to his wife and son for their support and patience.

TABLE OF CONTENTS

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

SILENCE AND VOICED-UNVOICED-MIXED EXCITATION CLASSIFICATION OF SPEECH WITH APPLICATIONS: A TWO-CHANNEL AND A ONE-CHANNEL

BY

Minsoo Hahn

December, 1989

Chairman: Dr. Donald G. Childers
Major Department: Electrical Engineering

In this study, we present two different algorithms for automatically classifying speech into four categories: silent and speech produced by three different excitation modes, i.e., voiced, unvoiced, and mixed (a combination of voiced and unvoiced). The algorithms employ information from two-channels (speech and EGG) and one-channel (speech only). Both algorithms were tested on the same data from six speakers, three male and three female, each speaking five sentences. An overall correct classification rate of 98.7% was achieved for the two-channel algorithm, when judged against skilled manual classification. This is superior to previously reported schemes. For the one-channel algorithm, the overall correct rate was slightly less, at 96.9%. This rate is still good enough to recommend the use of the algorithm in practical situations.

Simple modifications were made on the four-way classification algorithms in order to generate endpoint information and codewords. The modified algorithms were tested on a data set with sixty words, the digits from "one" to "ten" spoken by three male and three female speakers. Results showed that the algorithms would work reasonably well for endpoint detection, but when codeword generation is concerned, they need some additional smoothing filters.

Finally, after comparison of vocal fold opening intervals of voiced and mixed sounds, the suggestion was made to use a 25% longer glottal excitation waveform in the high quality synthesis of mixed sounds.

CHAPTER 1
INTRODUCTION

1.1 Research Rationale

Segmentation of speech according to its excitation mode into the categories of voiced, unvoiced, silence, and perhaps mixed (henceforth: V, U, S, and M) is required in many areas of speech processing and coding such as speech interpolation, vocoding, and speech recognition. The accuracy of segmentation is one of the important factors that affects the overall system performance directly. A variety of approaches has been described in the speech literature for accomplishing this acoustic segmentation [1–13].

It is well known that in a two-way telephone conversation, speech activity occurs only about 40 percent of time [14]. Accordingly, the use of speech interpolation in long distance telephony can double channel capacity without increasing the facilities of the transmission medium. Another possible application is the transmission of different information, such as printed text, graphs, and digital images, along with the speech signal during the silent intervals.

In the most commonly used model of speech production, whether it is a formant or a linear predictive coding vocoder, the speech signal is decomposed into two components [2,4,15,16,17,18]. One is a filter component representing the human vocal tract and the other is an excitation

component imitating human vocal fold vibration or the turbulent air flow needed for unvoiced sound production. In Figure 1-1, the block diagram of a typical speech synthesizer is shown [18]. In general, the excitation is represented by one of two states, voiced or unvoiced. Using this type of model, considerable success has been achieved by employing pattern classification techniques to assign a segment of speech to one of the two classes, voiced or unvoiced [4,5,6,8,9,11,13].

Despite the widespread use of this simplified model, the restriction of the excitation to the two classes is not adequate for the high quality speech synthesis from analysis parameters. Experiments show that high quality speech synthesis requires mixed excitation for synthesis of the voiced fricatives (such as /v/ in 'van', /ð/ in 'those', /z/ in 'zany', and /3/ in 'azure'). The pronunciation of such sounds requires the vibration of the vocal cords in conjunction with a turbulent air flow at some point of constriction in the vocal tract.

Synthesizers driven from stored data rather than from analysis parameters commonly include a link between the unvoiced and the voiced excitation in the synthesized speech. In order to allow a mixed source in an analysis-synthesis system, the excitation for a segment of speech must be identified as voiced, unvoiced, or mixed, i.e., a combination of voiced and unvoiced.

Current topics in the third major area of speech processing, speech recognition research, include 1) isolated word recognition (IWR), 2) continuous speech recognition, 3) speaker identification, and 4) speech understanding. Among these, the IWR system of large vocabularies has

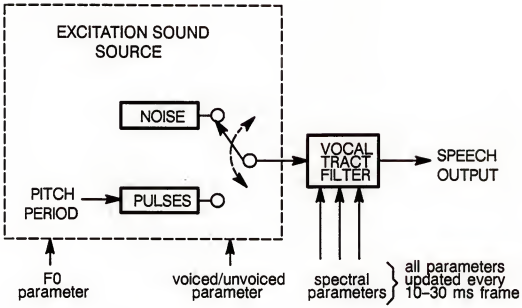


Figure 1–1. Basic Electrical Model of Speech Production [18]

recently received increased attention, because most current continuous speech recognition systems and speech understanding systems adopt the "word" as their template unit. As a result, improvements in IWR systems will directly affect the performance of these systems.

Unfortunately, many current IWR systems can not be extended to handle vocabularies of more than a few hundred words. This is partly due to the unavoidable choice between high hardware costs and unacceptably slow response times when one attempts to recognize an utterance by searching a large vocabulary exhaustively using template-matching techniques alone. To solve this problem, various strategies have been tried. Among them, the two most promising approaches are phoneme-based and two-pass techniques.

Phoneme-based IWR systems started with the idea that any spoken American utterance can be represented successfully with about 40 phonemes [2,19,20,21,22]. In these systems, reference word templates are stored as the phonemic transcription of the words. For example, the word 'level' is stored as 'levl' in its template. When the system meets a new input utterance, the phonemic classifier of the system extracts the phonemic transcription of an input utterance before pattern matching is executed. As we can easily see, the hardest step in this kind of IWR system is to realize a reliable phonemic classifier. Most of the effort has been concentrated on this phonemic classifier in order to improve the overall performance of the IWR system, but still no satisfactory result has been reported [22,23-26].

The two-pass IWR system generally adopts an acoustic segmentizer and a stress analyzer to get a "codeword" of the input utterance before

detailed pattern matching is executed [27,28,29,30]. This codeword plays the important role of reducing the number of possible word candidates among whole vocabularies of the system. For example, if the acoustic segmentizer and the stress analyzer of the system are reliable, the six words "ample", "apple", "natural", "echo", "neutral", and "ankle" can all be represented as one codeword: [stressed-voiced] [silence] [unvoiced] [voiced]. The codeword for this type of IWR system is usually in the form of linear prediction coefficients or formant information.

1.2 Literature Survey

There have been many studies on acoustic segmentation of speech, with results ranging from simple speech detection algorithms to voiced-unvoiced-mixed-silence classification algorithms. Some of them are listed in Table 1-1, and three important works are selected and described briefly below. (Henceforth: "A-B" for the separation into categories A and B, and "A/B" for a combined category consisting of A and B.)

1.2.1 Atal and Rabiner's V-U-S Classifier [4]

The input data for this classifier were two sentences, 1) "Should we chase those young outlaw cowboys?" and 2) "Few thieves are never sent to the jug." The features for the classification algorithm are, 1) zero crossing rate, 2) speech energy, 3) the correlation between adjacent speech samples, 4) the first predictor coefficient from a 12 order linear predictive coding analysis, and 5) the energy of the linear prediction error signal. A final correct classification rate of 96.6% was reported. Unfortunately,

Table 1-1. Studies on the acoustic segmentation of speech

RESEARCHER	YEAR	TYPE	WHERE	CORRECT RATE(%)
ATAL & RABINER [4]	1976	V-U-S	AT&T	96.6
RABINER et al. [5]	1977	V-U-S	AT&T	95.0
DAABOUL & ADOUL [6]	1977	V-U-S	SH. U.	95.0
SIEGEL & BESSEY [7]	1982	V-U-M	PURDUE	94.0
LARAR [10]	1985	V-U-M-S	UF	95.5

SH. U. in this table is for Sherwood Unversity.

there was no comment on how the voiced fricatives in the input sentences were handled and the data set seemed to be too small to produce a generalizable result.

1.2.2 Siegel and Bessey's V-U-M Classifier [7]

This system used eight sentences as its input: 1) "Why did the measuring jar sink fast?" 2) "In Chapter Six, we had better discuss the passing of the old codger," 3) "Forget your rotten games of pleasure," 4) "The mute soldier gazes up," 5) "These three machines talked with cabbage plants," 6) "The thin prisoner chases the fat judges," 7) "Then Cassie visited the Grand Rapids Jail," and 8) "Bill has not stopped me yet, although he should have." Among these, sentences 1, 2, 3, 4, and 5 were used as a training set and the rest were used only for the test. The features adopted were 1) speech energy, 2) normalized autocorrelation coefficient at unit sample delay, 3) linear prediction error, 4) the first linear predictor coefficient, 5) zero crossing rate, 6) the ratio of energy in the signal above r(H) Hz to that below r(L) Hz (in fact, three ratios were used), and four others. The final recognition rate of 94.0% was asserted for one of the data sets.

There are four major disadvantages to this classifier: 1) It accepts only the speech part of input sequences, so an operator was needed to eliminate silent intervals from the input sentences manually before they were processed. As a result, the performance of the system was greatly improved, but it made the system less attractive due to the incapability of providing endpoint information essential to isolated word recognition (IWR)

systems. 2) The author failed to describe clearly how mixed intervals were identified manually. 3) The absence of silent interval detection capability can be critical in some applications, such as automatic control of the excitation mode in speech synthesis and codeword generation in IWR systems. 4) Lastly, the final overall error rate was not given.

1.2.3 Larar's V-U-M-S Classifier [10]

This classifier was a two-channel one, using both speech and EGG (electroglottography) as its input signals. This system was realized mainly for the purpose of improving the performance of a 100 word vocabulary IWR system by helping to select an acoustically equivalent subset based on codewords. The words "thirteen", "seven", "zero", "ten", "five", and "twelve" were tested on the system, yielding a 95.45% final recognition rate, while a 87.5% correct rate was achieved for mixed sound identification.

The features used were 1) the zero crossing rate of speech signal, 2) the energy of EGG signal, and 3) the energy of speech signal. The V/M-U/S classification was heavily dependent on the presence of the relatively high EGG energy in the frame. However, the use of the energy of the EGG signal in order to detect vocal fold vibration, can deteriorate the system significantly when there exists a relatively large low frequency fluctuation in the signal.

The data set can be seen to have two weak points: 1) the number of silent frames is too large, 69.8% of total frames, to declare that the 95.45% correct rate is a generalizable one, and 2) the number of mixed

frames is too small to assert the final mixed frame identification rate, 87.5%, as a reliable one.

1.3 Objective

Most existing acoustic segmentizers produce error rates of only about 5% whether they are three-way or four-way ones. But as shown above, the performance of an acoustic segmentizer directly affects the quality of synthesized speech and the performance of the two-pass IWR system. In this sense, existing acoustic segmentizers are far from being satisfactory, especially in case of mixed sound identification. In order to get a more satisfactory result with either a speech synthesizer or a two-pass IWR system, more research has to be concentrated on improving the overall performance of the acoustic segmentizer and more emphasis has to be given to the development of a more accurate classification algorithm for mixed sounds.

The main objective of this study is to design a more reliable acoustic segmentizer, capable of segmentizing input utterances into the four categories of voiced, unvoiced, mixed, and silence. Two approaches are explored.. One is a two-channel four-way classification algorithm using both the speech and the EGG (electroglottography) signals and the other is a one-channel classification algorithm using speech signal only. Both approaches are utilized to accomplish the accurate endpoint detection essential to time registration (or alignment) of input utterance and template in an IWR system. Another application of codeword generation is tested for future use in the two-pass IWR systems. The chapters that follow

describe the design, implementation, and testing of a minicomputer-based laboratory realization.

1.4 Description of Chapters

The techniques associated with reference data collection and processing are discussed in Chapter 2. In Chapter 3, the design of a two-channel (speech and EGG) four-way (V-U-M-S) classifier is described and its performance is evaluated. In Chapter 4, a one-channel (speech only) four-way classification algorithm is explained and its performance is compared with that of the two-channel four-way classifier. In Chapter 5, some applications of the two classifiers in a two-pass IWR system are examined, such as endpoint detection and codeword generation with a 10 digit vocabulary. A new glottal excitation model for the mixed sound production is also suggested in this chapter. Chapter 6 is devoted to the concluding remarks, including an indication of areas where future endeavors may prove fruitful.

CHAPTER 2
DATA COLLECTION AND PREPROCESSING

2.1 Data Collection

2.1.1 Description of the Computer System

The data collection system used for this research is shown in Figure 2-1. An Electro-Voice RE-10 microphone was used to convert the acoustical pressure of the speech sound into an electrical signal. This microphone has a very good frequency response at frequencies above 50 Hz, but cuts off the low-frequency component below 50 Hz. A Synchrovoice Inc. electroglottographic detector was selected to collect the EGG signal. Details of its working principle appear in the next section. Two Digital Sound Corporation model DSC240 preamplifiers enabled us to record and to replay both the speech and EGG signals, and a DSC-200 digitizer multiplexed the speech and EGG signals synchronously, with a sampling frequency of 20 kHz and resolution of 16 bits per sample. Finally, a VAX 11/750 computer system managed all of the processing procedures, such as activating the preamplifiers and the digitizer, storing the collected data, and replaying the digitized speech.

During data collection, in order to confirm the validity of input data, an oscilloscope monitored both the speech and EGG signals to show that both signals were properly amplified. A loudspeaker reproduced the

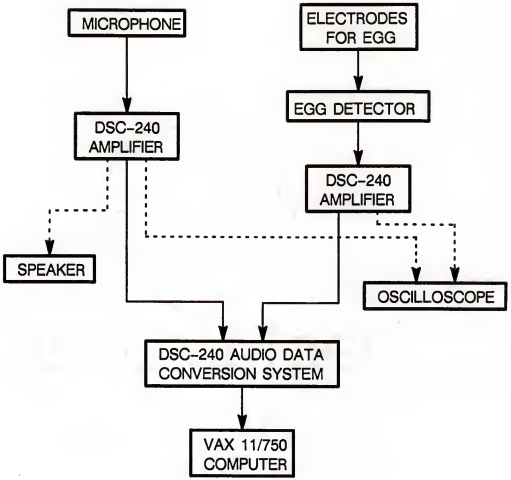


Figure 2-1. Data collection system

digitized speech with the help of a DSC–240 preamplifier, and thus, provided information about the quality of the digitized speech.

2.1.2 Electroglottography Detection

Electroglottography (EGG) is based on the electrical transmission of a high–frequency current through the tissues at the glottal levels. A weak alternating current, in the order of microampere, is applied to electrodes which are in direct contact with the skin of the neck on each side of the larynx. A signal generator, which may be either of constant voltage or constant current type, activates these electrodes. The frequency of the activating signal is usually in the magnitude of several MHz and the voltage level is about 0.5 volt, depending on the tissue impedance and current.

The vibrating vocal folds constitute a varying impedance path that modulates a small part of the radio frequency current transmitted between the two electrodes. These modulations can be detected and amplified to obtain the EGG signal. A functional block diagram of an electroglottograph with a typical EGG signal is shown in Fig 2–2. The change in impedance across the larynx is primarily due to the change in the lateral contact area of the vocal folds [31,32,33]. Hence most speech researchers believe that the EGG is a measure of the amount of the vocal folds' contact area, but not of the area of the glottis. However, it has been impossible so far to confirm this by impedance measurement, and the factors causing the depicted changes in laryngeal impedance are not known in detail.

In order to be useful to speech researchers, the EGG waveform must be related to the vocal fold vibration cycle. It can best be understood by

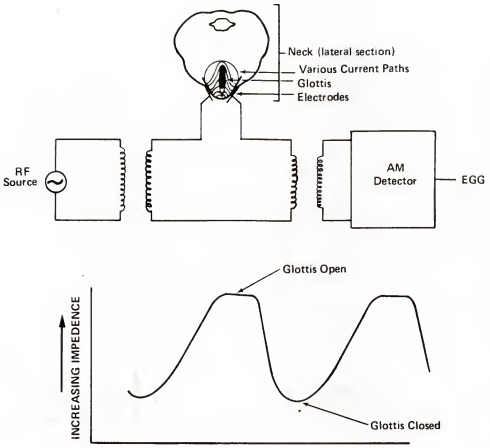


Figure 2-2. A system configuration for the electroglottography (a) and the output EGG waveform (b)

comparing the EGG signal with the glottal area function. Many researchers explored this relationship using the EGG signal with synchronized stroboscopy and ultra-high-speed cinematography along with the glottal waveform [34,35]. Childers et al. [34] found that the point of maximum negative value in a differentiated EGG signal agrees well with the closing time of the vocal folds. However, as shown in Figure 2-2, the open phase of the EGG signal normally lacks details, as the impedance is equally maximum whether the glottal area is narrow or wide. Therefore, it should be noted that to find the point of maximum glottal opening with the aid of the EGG signal is impossible with current techniques.

Although the EGG signal does not seem appropriate for detailed monitoring of the glottal vibration cycle, its simple configuration with one steep deflection in every period makes it ideally suitable for measurement of the pitch period that is inversely related to the fundamental frequency of the voice. That is why the EGG signal has been used frequently for the reliable measurement of the fundamental frequency of speech.

The Mind-Machine Interaction Research Center at the University of Florida has conducted extensive studies on the EGG signal with synchronized high-speed film data and speech signal [34]. Krishnarmurthy and Childers [1] developed a pitch synchronous formant tracking algorithm with the aid of EGG signal and suggested a possible use of EGG signal in voiced-unvoiced discrimination of speech. As described in Chapter 1, Larar [10] developed a voiced-unvoiced-mixed-silence speech signal discrimination algorithm with a moderate correct rate. His work relied heavily on the EGG signal as a strong indicator of vocal fold vibration.

Many other studies related to speech pathology have also been carried out, such as Alsaka's [36] and Bae's [37].

2.1.3. Data Base for Four-way Classification

For the data base of the four-way classification algorithm in this study, five sentences were selected based on their phonetic contents. These five sentences are

Sentence 1: We were away a year ago.

Sentence 2: Early one morning a man and a woman ambled along a one mile lane.

Sentence 3: Should we chase those cowboys?

Sentence 4: That zany van is azure.

Sentence 5: We saw the ten pink fish.

Three male and three female speakers were asked to utter these five sentences one by one with comfortable speed, tone, and loudness in an IAC (Industrial Acoustics Company) sound booth. With six speakers and five sentences, the total number of sentences for the study was thirty. In terms of phonetics, Sentence 1 is composed of all voiced, vocalic sounds, Sentence 2 adds nasals and liquids. Sentence 3 adds fricatives and affricates, while Sentence 4 contains all the voiced fricatives of English. Finally, Sentence 5 has unvoiced fricatives and plosives. All the file names and their lengths, as stored in our computer system, are shown in Table 2-1. (File names for EGG data have the same names as corresponding speech data except that they have extensions beginning with 'e' instead of

Table 2-1. Description of data for four-way classification

Sentence (Sex)	File Name	Data Length
Sentence 1-a (M)	nraaan025.smst	18688
Sentence 1-b (M)	nrdrwn025.smst	18942
Sentence 1-c (M)	nrjrsn025.smst	20223
Sentence 1-d (F)	nrcxon025.sfst	19968
Sentence 1-e (F)	nrbemn025.sfst	19968
Sentence 1-f (F)	nrmbkn025.sfst	21760
Sentence 2-a (M)	nraaan026.smst	45056
Sentence 2-b (M)	nrdrwn026.smst	43520
Sentence 2-c (M)	nrjrsn026.smst	42496
Sentence 2-d (F)	nrcxon026.sfst	38656
Sentence 2-e (F)	nrbemn026.sfst	41472
Sentence 2-f (F)	nrmbkn026.sfst	43264
Sentence 3-a (M)	nraaan027.smst	18432
Sentence 3-b (M)	nrdrwn027.smst	20992
Sentence 3-c (M)	nrjrsn027.smst	20224
Sentence 3-d (F)	nrcxon027.sfst	19968
Sentence 3-e (F)	nrbemn027.sfst	20224
Sentence 3-f (F)	nrmbkn027.sfst	20224
Sentence 4-a (M)	nra1an001.smwt	21759
Sentence 4-b (M)	nrd1cn001.smwt	22528
Sentence 4-c (M)	nrd1hn001.smwt	21504
Sentence 4-d (F)	nrd1hn001.sfwt	26624
Sentence 4-e (F)	nrm1kn001.sfwt	20480
Sentence 4-f (F)	nrn1sn001.sfwt	21504
Sentence 5-a (M)	nra2an001.smwt	21504
Sentence 5-b (M)	nrd2cn001.smwt	24576
Sentence 5-c (M)	nrm1gn001.smwt	20992
Sentence 5-d (F)	nrd2hn001.sfwt	20480
Sentence 5-e (F)	nrm2kn001.sfwt	25600
Sentence 5-f (F)	nrb1cn001.sfwt	20736

's'.) For convenience, the sentences will be referred to with names like 'sentence 1-a' instead of by file names, such as 'nraaan025.smst'.

2.1.4 Data Base for Applications

For the study of speech recognition and synthesis applications of the four-way classification algorithm, a data base consisting of the digits from one to ten was used. The same number of speakers, i.e., three males and three females, pronounced these words discretely one by one under the same environmental conditions used for the collection of the data for four-way classification. At first, ten digits from a speaker were collected and stored in one data file. After this, each digit was extracted from the file with manual examination and confirmed by replaying it on a loudspeaker. With six speakers and ten digits, sixty data files were generated. Table 2-2 shows the entire file names and their contents in this data set. As before, names like 'word 1-a' are preferred for simplicity to names like 'the "one" from nraaan001.smwt'.

2.2 Preprocessing of Data

2.2.1 Demultiplexing and Trimming the Data

The collected data are two-channel (speech and EGG) multiplexed signals sampled at 20 kHz, and contain a large portion of silence at the beginning and end of each file. The signal is demultiplexed to produce both the digitized speech and EGG signals with the sampling frequency of 10 kHz. These demultiplexed signals are trimmed to get rid of unnecessary

Table 2-2. Description of data for applications

Word (Sex)	Source File	Data Length
Word 1-a (M)	nraaan001.smwt	6000
Word 1-b (M)	nrdrwn001.smwt	6000
Word 1-c (M)	nrjrsn001.smwt	5000
Word 1-d (F)	nrcxon001.sfwt	6800
Word 1-e (F)	nrbemn001.sfwt	5500
Word 1-f (F)	nrmbkn001.sfwt	6500
Word 2-a (M)	nraaan001.smwt	5500
Word 2-b (M)	nrdrwn001.smwt	6000
Word 2-c (M)	nrjrsn001.smwt	6000
Word 2-d (F)	nrcxon001.sfwt	6700
Word 2-e (F)	nrbemn001.sfwt	6500
Word 2-f (F)	nrmbkn001.sfwt	7000
Word 3-a (M)	nraaan001.smwt	6000
Word 3-b (M)	nrdrwn001.smwt	6000
Word 3-c (M)	nrjrsn001.smwt	6000
Word 3-d (F)	nrcxon001.sfwt	6500
Word 3-e (F)	nrbemn001.sfwt	6000
Word 3-f (F)	nrmbkn001.sfwt	6000
Word 4-a (M)	nraaan001.smwt	7000
Word 4-b (M)	nrdrwn001.smwt	6000
Word 4-c (M)	nrjrsn001.smwt	5500
Word 4-d (F)	nrcxon001.sfwt	6000
Word 4-e (F)	nrbemn001.sfwt	6500
Word 4-f (F)	nrmbkn001.sfwt	7000
Word 5-a (M)	nraaan001.smwt	8000
Word 5-b (M)	nrdrwn001.smwt	6000
Word 5-c (M)	nrjrsn001.smwt	8000
Word 5-d (F)	nrcxon001.sfwt	7500
Word 5-e (F)	nrbemn001.sfwt	6500
Word 5-f (F)	nrmbkn001.sfwt	6500

Table 2–2.––Continued.

Word (Sex)	Source File	Data Length
Word 6–a (M)	nraaan001.smwt	7000
Word 6–b (M)	nrdrwn001.smwt	8000
Word 6–c (M)	nrjrsn001.smwt	7000
Word 6–d (F)	nrcxon001.sfwt	5500
Word 6–e (F)	nrbemn001.sfwt	9500
Word 6–f (F)	nrmbkn001.sfwt	8000
Word 7–a (M)	nraaan001.smwt	8000
Word 7–b (M)	nrdrwn001.smwt	6000
Word 7–c (M)	nrjrsn001.smwt	7000
Word 7–d (F)	nrcxon001.sfwt	8000
Word 7–e (F)	nrbemn001.sfwt	12000
Word 7–f (F)	nrmbkn001.sfwt	9000
Word 8–a (M)	nraaan001.smwt	6000
Word 8–b (M)	nrdrwn001.smwt	6000
Word 8–c (M)	nrjrsn001.smwt	8000
Word 8–d (F)	nrcxon001.sfwt	5000
Word 8–e (F)	nrbemn001.sfwt	6500
Word 8–f (F)	nrmbkn001.sfwt	7000
Word 9–a (M)	nraaan001.smwt	6000
Word 9–b (M)	nrdrwn001.smwt	6000
Word 9–c (M)	nrjrsn001.smwt	5000
Word 9–d (F)	nrcxon001.sfwt	7000
Word 9–e (F)	nrbemn001.sfwt	6500
Word 9–f (F)	nrmbkn001.sfwt	7000
Word 10–a (M)	nraaan001.smwt	5000
Word 10–b (M)	nrdrwn001.smwt	7500
Word 10–c (M)	nrjrsn001.smwt	5000
Word 10–d (F)	nrcxon001.sfwt	6000
Word 10–e (F)	nrbemn001.sfwt	5500
Word 10–f (F)	nrmbkn001.sfwt	6000

surplus silent data at the beginning and end of each utterance to save computer memory. While trimming, the operator left at least five silent frames at the beginning of each utterance. These frames would be used to obtain the statistics for silence such as the average zero crossing rate and the average energy level, which are essential to the four-way classification algorithm.

2.2.2 Synchronization of Data

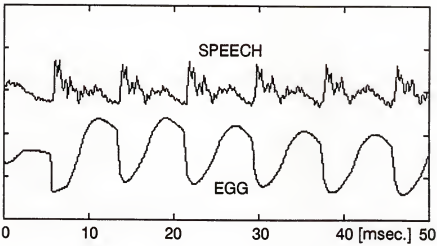
The microphone was kept 6 inches (15.24 centimeters) away from the speaker's lips to reduce breath noises and to simplify the alignment procedure. Synchronization of the speech and EGG waveforms is necessary to account for the time delay while the speech signal travels from the vocal folds to the microphone. This time delay can be expressed as follows.

$$T_d = (VT_l / C_{VT}) + (SM_l / C_{air}) \qquad (2.1)$$

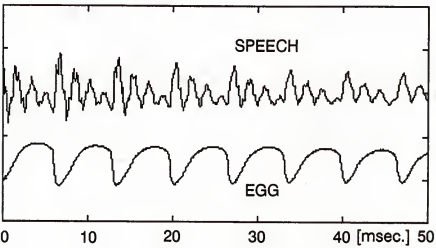
where T_d is the time delay in seconds and VT_l is the vocal tract length in centimeters. The distance from the speaker's lips to the microphone, 15.24 centimeters in this study, is denoted as SM_l. C_{VT} and C_{air} are for the velocities of sound in the vocal tract and in air, respectively. If we select typical values of these parameters, e.g., VT_l of 17.0 centimeters (for adult male subjects), C_{VT} of 35300 cm/sec [20,38], and C_{air} of 34400 cm/sec, the T_d obtained is 0.925 milliseconds. Hence the number of data points to be discarded from the beginning of the speech record is nine.

The matter of variation in vocal tract lengths among adult males was largely resolved with the 17.0 centimeter compromise. Equation (2.1) shows that a nine-data-point correction is actually appropriate for vocal tract

lengths from 14.4 to 17.9 centimeters long. On the other hand, the average length of the vocal tracts among adult females is known to be 14.0 centimeters [19], and this leads to a one-data-point misalignment of the speech and EGG signals. This misalignment does not cause any serious problem in the design of a reliable four-way classification algorithm because a segment size of 100 data points would be used. Examination of the data also supported the use of this nine-data-point correction for adult speakers. Examples of aligned speech and EGG signals for a male and a female speaker are shown in Figure 2-3. Both speech signals in this figure came from the /e/ sound in word "ten".



(a)



(b)

Figure 2–3. Synchronized speech and EGG signal: (a) male and (b) female

CHAPTER 3
TWO-CHANNEL FOUR-WAY CLASSIFICATION

3.1 Introduction

Two-channel four-way classification is the acoustic classification of segments of the speech signal into the four excitation categories of voiced, unvoiced, mixed, or silent, with the use of both the speech signal and the EGG signal. Voiced sounds are speech sounds pronounced with vocal fold vibration and with, as a result, a speech waveform showing quasi-periodic characteristics. In American English, all vowels and certain consonants, like /b/, /g/, /l/, and /r/, belong to voiced sounds [19,20,21]. Unvoiced sounds are uttered without vocal fold vibration, but with a constriction in the vocal tract which produces a turbulent air flow and, as a result, generates a noise-like speech waveform. Examples of unvoiced sounds are /f/, /s/, /p/, /t/, and /k/. (The plosive releases of /p/, /t/, and /k/ are preceded by silence.) Mixed sounds can be considered as a combination of voiced and unvoiced sounds. Namely, they are generated with both vocal fold vibration and a vocal tract constriction causing a turbulent air flow, which makes the speech waveform look like noise with a low frequency carrier. The phonemes, /v/ and /z/ are typical examples of mixed sounds in American English. Lastly, silence can be defined as either a pause in speech or background noise.

In Figure 3-1, speech and EGG waveforms for each type of sound are presented. Each example is 20 milliseconds long.

In order to label each speech segment according to its excitation mode, an analysis frame size of 10 milliseconds was selected, which amounts to 100 data points at a 10 kHz sampling frequency. The features selected for use in the classification algorithm are

1) the energy of the speech signal,
2) the zero crossing rate of the speech signal,
3) the level crossing rate of the speech signal,
4) the zero crossing rate of the differentiated speech signal, and
5) the level crossing rate of the differentiated and normalized EGG signal.

In Figure 3-2, the speech, EGG, differentiated speech, and differentiated and normalized EGG signals are shown as an illustration. The illustration comes from sentence 4-d and is part of 'azure' with some additional silent frames at the beginning. Even though the phonemic transcription of this part will produce only voiced, mixed, and (added) silent intervals, a careful manual inspection shows that all four categories of voiced, unvoiced, mixed, and silent exist in this part. This is not unusual because some people pronounce a mixed sound in the normal way with vocal fold vibration, some utter it as mixed at the beginning of the phoneme but devoiced for the latter part, and the rest pronounce it as completely unvoiced [5,39]. In Figure 3-2-a, boundaries obtained by a manual classification are depicted.

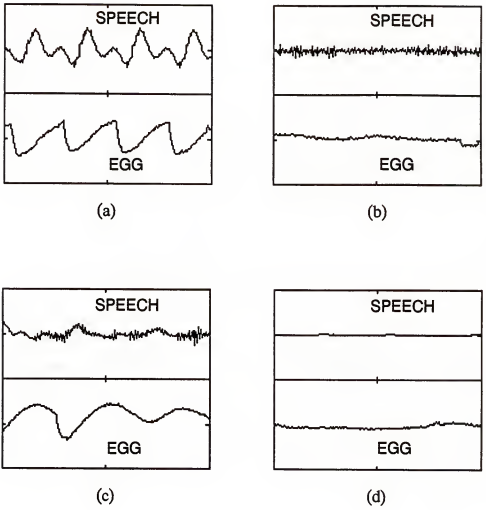


Figure 3–1. Examples of speech and EGG waveform: (a) voiced, (b) unvoiced, (c) mixed, and (d) silence

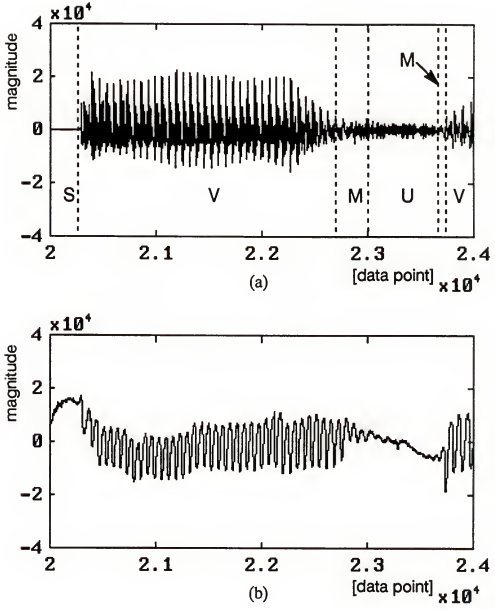


Figure 3-2. Examples of signals: (a) speech, (b) EGG

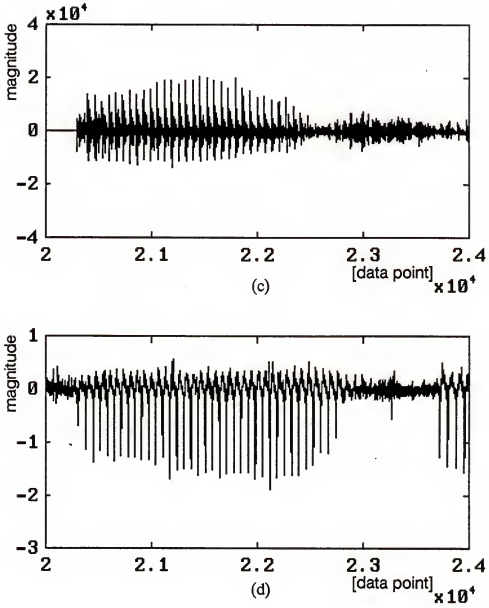


Figure 3–2.––Continued: (c) differentiated speech, (d) differentiated and normalized EGG

Six sentences, three from male and three from female speakers (sentence 1–a, sentence 1–d, sentence 3–a, sentence 3–e, sentence 4–a, and sentence 4–e), were used as a training set for this two–channel four–way classifier, and the final classification algorithm was applied to all the thirty sentences to evaluate its overall performance.

3.2 Algorithmic Details

The algorithm can be divided into three main parts as shown in Figure 3–3. The five basic features are calculated for every frame of the input sentence and are used for an early classification of the frames that are clear cases of voiced and unvoiced. Statistics, such as averages and standard deviations, are calculated using the five features of these clear–cut frames, for use directly in the tree–structure pattern classification algorithm, which follows. In that step, the remaining, more difficult input speech segments are assigned to all four categories of voiced, unvoiced, mixed, or silent, according to a tree–structure pattern classification technique using the five features and their statistics. The last step of the algorithm is the error correction step utilizing general acoustic characteristics of human speech. For example, errors such as VVVUVVV and SSSUSSS are corrected to VVVVVVV and SSSSSSS.

3.2.1 Feature Extraction

Unfortunately, there is no general rule about which features to use for the best result in a given task, and feature selection is heavily dependent on the algorithm designer's experience or insight into the objects

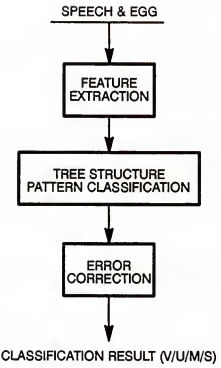


Figure 3-3. Block diagram of two-channel four-way classifier

to be classified [40,41]. (It might even be said that the quality of features is as good as that of the algorithm designer.)

For the two-channel four-way classification algorithm, the five time-domain features listed above were selected. The basic underlying reasons for choosing these features are as follows:

1) The energy of the speech signal can be a cue for silence-speech and voiced-unvoiced classifications. In general, it has a larger value for voiced than for unvoiced sounds and has a smaller value for silent than for any combination of voiced, unvoiced, and mixed intervals.

2) The level crossing rate of the differentiated and normalized EGG signal is a strong indicator of the existence of vocal fold vibration and helps in the voiced/mixed-unvoiced/silent classification algorithm.

3) The zero crossing rate of the speech signal has a small value for silent, a large value for unvoiced, and an intermediate value for voiced or mixed frames, and it helps in the silence-speech and unvoiced-voiced/mixed classifications.

4) The level crossing rate of the speech signal has a relatively large value for speech but has a small value, near zero, for silence, and mainly helps in the silence-speech classification.

5) The zero crossing rate of the differentiated speech signal has a large value for unvoiced or mixed frames and can be a cue for unvoiced/mixed-voiced/silent classification. This feature is selected to prevent a possible error in detecting the noise-like property of

mixed sound. The zero crossing rate and the level crossing rate of the speech signal often fail to give this information because of the low frequency (carrier) component in mixed sounds.

For the first part of the feature extraction step, all five basic features were calculated for every frame of the unique utterance. As an illustration, the five features were evaluated for the same utterances as in Figure 3-2, and the results are presented in Figure 3-4 as an example. By comparing these features and the original speech signal (given in Figure 3-2 with manually classified boundaries) one can get a rough idea about the relationship between the features and each type of sound.

After the five features were calculated, an "early" classification designates certain frames as "clear-cut" instances of voiced or unvoiced speech. This early classification was based on two decision rules using the five features. Specifically, if a frame had ELCR(i) greater than 0.7 and SLCR(i) less than 20, it was considered a clear-cut voiced frame, and if a frame had ELCR(i) less than 0.05, SENG(i) less than the sum of SEAV and 5SESIG, and SDZCR(i) greater than the sum of VDZAV and 3VDZSIG, it was considered a clear-cut unvoiced frame. (These parameters are explained below.)

At this point in the feature extraction step, frames may have been assigned to three categories: voiced and unvoiced from the early classification, and silent from the five frames at the beginning of the input data. As the final step of feature extraction, the statistical characteristics were calculated for each of the three categories, e.g., the average and standard deviation of the zero crossing rate for clear-cut unvoiced sounds.

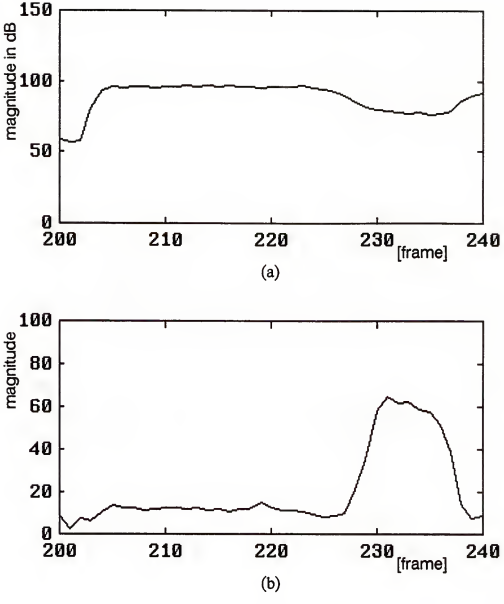


Figure 3–4. Examples of Features: (a) speech energy, (b) zero crossing rate of speech

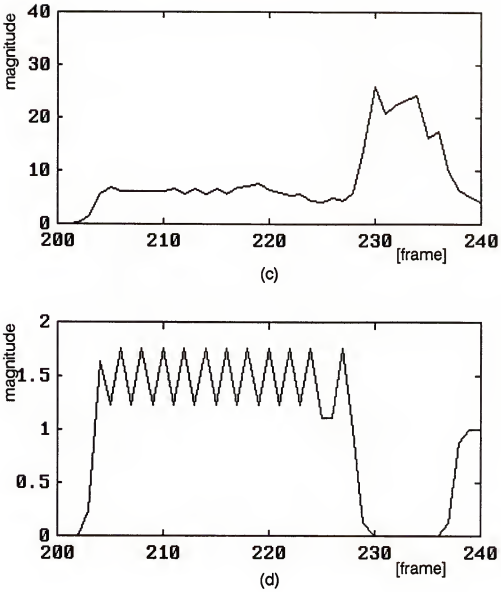


Figure 3-4.--Continued: (c) level crossing rate of speech
(d) level crossing rate of EGG

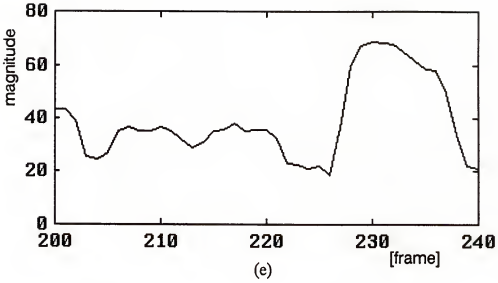


Figure 3–4.—Continued: (e) zero crossing rate of differentiated speech

In Figure 3-5, the details of the feature extraction step are shown. It must be remembered in going through this step that statistics are only evaluated for the clear-cut voiced and unvoiced frames (and five beginning silent frames), rather than for a preclassified training sentence. The merit of this strategy, as opposed to the latter, is that our algorithm will have the capability of adaptation to the properties of any input sentence by changing its threshold values automatically. (Most algorithms reviewed in Chapter 1 were not adaptive ones and could produce unacceptable results when a strange speaker was subjected.)

In order to understand the details of this feature extraction step, explanation of the parameters is essential. In the following definitions the index 'i' was used for frames, the index 'k' was used for data points across frames, while 'j' was used for data points within a frame.

SCH(k): the k-th data point of the speech signal.

EGG(k): the k-th data point of the EGG signal.

DSCH(k): the k-th data point of the differentiated speech signal.

$$DSCH(k) = SCH(k) - SCH(k-1) \tag{3-1}$$

DEGG(k): the k-th data point of the differentiated EGG signal.

$$DEGG(k) = EGG(k) - EGG(k-1) \tag{3-2}$$

DNEGG(k): the k-th data point of the differentiated and normalize EGG signal.

$$DNEGG(k) = DEGG(k)/MAXDEGG \tag{3-3}$$

where MAXDEGG is the maximum value of the rectified DEGG signal in a sentence.

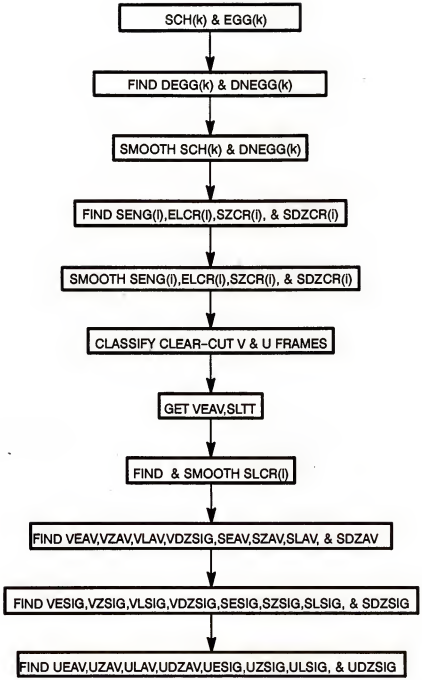


Figure 3-5. Feature extraction step (Two-channel)

SENG(i): the energy in decibel (dB) of the i-th frame of the speech signal.

$$SENG(i) = 10*LOG(\epsilon + \sum_{j=1}^{100} SCH(i*100+j)^2) \qquad (3\text{–}4)$$

where ϵ is a small positive constant added to prevent the computing of log of zero. In this study, ϵ was set to 0.0001.

SZCR(i): the zero crossing rate of the i-th frame of the speech signal. The value of SZCR(i) is incremented by one when the product of SCH(i*100+j) and SCH(i*100+j-1) is less than zero.

SDZCR(i): the zero crossing rate of the differentiated speech signal.

ELCR(i): the level crossing rate calculated for the i-th frame of the differentiated and normalized EGG signal, where ELCR(i) is incremented by one if DNEGG(i*100+j-1) is greater than -0.5 and both DNEGG(i*100+j) and DNEGG(i*100+j+1) are less than -0.5. Here, a three-point level crossing detector is applied because some EGG signals have a relatively large noise level and produce a false level crossing when a two-point detector is used.

SLCR(i): the level crossing rate of the i-th frame of the speech signal. The value of SLCR(i) is incremented by one when the product of (SCH(i+100+j)-SLTT) and (SCH(i*100+j-1)-SLTT) is less than zero.

IVUS(i): the classification result of the i-th frame represented in number. The values 1, 4, 7, and 9 are arbitrarily assigned to silent, unvoiced, mixed, and voiced, respectively.

SLTT: The threshold value to calculate the level crossing rate of the speech signal. It was set to 10% of the average magnitude of rectified voiced/mixed sounds.

VEAV: the average energy of voiced sounds.

VESIG: the standard deviation of the energy of voiced sounds.

VZAV: the average zero crossing rate of voiced sounds.

VZSIG: the standard deviation of the zero crossing rate of voiced sounds.

VLAV: the average level crossing rate of voiced sounds.

VLSIG: the standard deviation of the level crossing rate for voiced sounds.

VDZAV: the average zero crossing rate of differentiated voiced sounds.

VDZSIG: the standard deviation of the zero crossing rate for differentiated voiced sounds.

The eight statistics above (with variable names in "V") were all calculated based on the frames classified as clear–cut voiced frames in the "early" classification. There are analogous statistics for the clear–cut unvoiced frames (with variable names in "U"), and for the five silent frames at the beginning of each utterance (with variable names in "S"). All these statistical values were calculated on a sentence by sentence basis, using the clear–cut voiced and unvoiced frames and the five silent frames in the sentence, and can therefore be considered as adaptive statistical values.

In smoothing SCH(k), DNEGG(k), SENG(i), ELCR(i), SZCR(i), SDZCR(i), and SLCR(i), a three-point filter of (0.12, 0.76, 0.12) was used. For example:

$$SCH(k) = 0.12*\{SCH(k-1)+SCH(k+1)\}+0.76*SCH(k) \quad (3-5)$$

This filter has linear phase characteristics and plays a similar role to a low pass filter.

All these features and statistics are used to determine threshold values for the second step, pattern classification. As an illustration, the statistics for the six training sentences are given in Table 3-1.

3.2.2 Pattern Classification

Figure 3-6 shows the details of the tree-structure pattern classification algorithm for the two-channel four-way classifier. There are some threshold values and rules based on the statistics obtained in the previous feature extraction step.

3.2.2.1 Threshold Explanation

The threshold values used in pattern classification algorithm are determined via two different methods according to whether the feature extraction step found any clear-cut unvoiced frame. When there is no clear-cut unvoiced frame in the input sentence, the threshold values are defined as follows.

$$ETH1 = (SEAV+VEAV-2*(SESIG+VESIG))/2 \quad (3-6-a)$$

$$ETH2 = (SEAV+VEAV)/2 \quad (3-7-a)$$

$$ETH3 = (SEAV+VEAV-SESIG-VESIG)/2 \quad (3-8-a)$$

Table 3-1. Statistical values (Two-channel)

FEATURES \ SENTENCE	1-a	1-d	3-a	3-e	4-a	4-e
SEAV (dB)	47.0	62.6	54.4	57.6	57.2	60.6
(SESIG)	(0.6)	(1.8)	(1.6)	(0.9)	(2.7)	(1.7)
VEAV (dB)	84.1	91.0	87.0	91.8	94.0	88.7
(VESIG)	(5.3)	(3.7)	(6.6)	(2.2)	(5.7)	(5.5)
UEAV (dB)	*	*	69.3	71.1	*	*
(UESIG)	(*)	(*)	(3.8)	(4.8)	(*)	(*)
SZAV	1.8	3.8	6.0	11.9	3.6	3.5
(SZSIG)	(0.8)	(1.9)	(7.0)	(4.8)	(1.8)	(1.6)
VZAV	8.5	8.4	10.9	9.9	10.8	10.2
(VZSIG)	(2.9)	(2.4)	(6.5)	(3.3)	(6.1)	(7.4)
UZAV	*	*	61.2	47.0	*	*
(UZSIG)	(*)	(*)	(12.3)	(20.7)	(*)	(*)
SLAV	0.0	0.0	0.0	0.0	0.0	0.0
(SLSIG)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
VLAV	4.3	4.2	5.2	5.0	5.3	4.8
(VLSIG)	(1.6)	(1.2)	(2.5)	(1.5)	(3.1)	(3.1)
ULAV	*	*	19.2	12.9	*	*
(ULSIG)	(*)	(*)	(8.7)	(9.3)	(*)	(*)
SDZAV	40.1	43.2	42.1	38.5	43.9	41.4
(SDZSIG)	(3.1)	(3.0)	(5.7)	(2.7)	(3.3)	(1.3)
VDZAV	21.9	17.5	25.5	19.9	28.9	28.6
(VDZSIG)	(7.4)	(7.0)	(12.0)	(7.9)	(13.2)	(10.9)
UDZAV	*	*	72.0	67.3	*	*
(UDZSIG)	(*)	(*)	(7.6)	(13.4)	(*)	(*)

* Data unavailable due to the absence of clear-cut unvoiced frames.

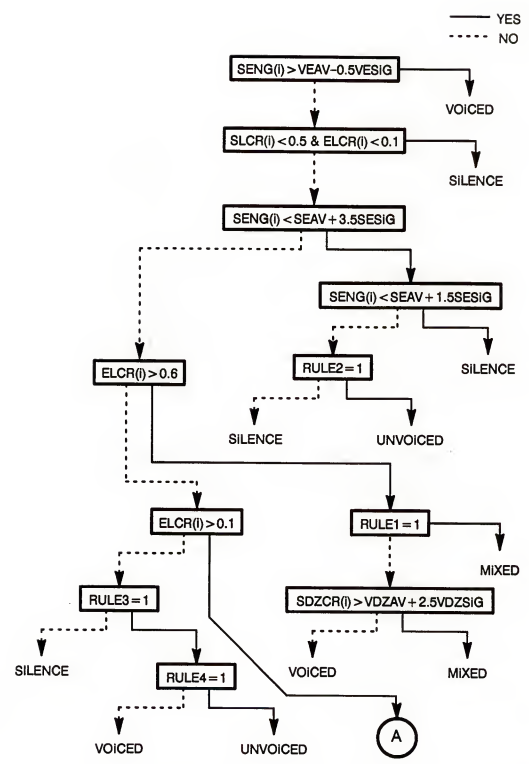


Figure 3-6. Details of pattern classification step (Two-channel)

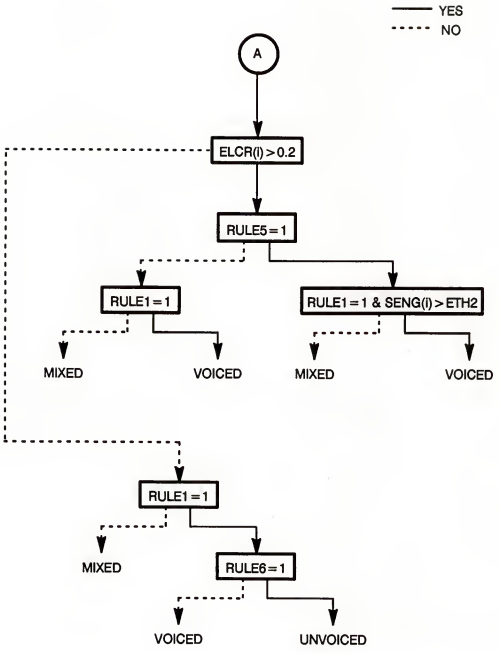


Figure 3-6.—Continued

If there are some clear-cut unvoiced frames in the sentence, the values are defined as follows.

ETH1 = UEAV-1.5*UESIG (3-6-b)

ETH2 = UEAV+UESIG (3-7-b)

ETH3 = UEAV (3-8-b)

By using the threshold and the features (with their statistics), some decision rules are determined. These rules appearing in Figure 3-6 are

RULE1 = 1, if SDZCR(i) is greater than VDZAV+VDZSIG, SZCR(i) is greater than VZAV, SLCR(i) is greater than VLAV-0.7*VLSIG, SDZCR(i) is greater than 45, and SENG(i) is greater than ETH3

= 0, otherwise

RULE2 = 1, if SLCR(i) is greater than 0.7 and SZCR(i) is greater than SZAV+2*SZSIG

= 0, otherwise

RULE3 = 1, if SLCR(i) is less than 0.7, SZCR(i) is less than SZAV+1.5*SZSIG, SENG(i) is less than ETH1, and SENG(i) is less than SEAV+4*SESIG

= 0, otherwise

RULE4 = 1, if SDZCR(i) is less than VDZAV+VDZSIG and SZCR(i) is less than VZAV+1.5*VZSIG

= 0, otherwise

RULE5 = 1, if SENG(i) is greater than ETH1, SLCR(i) is less than VLAV, and SZCR(i) is less than VZAV+3*VZSIG

= 0, otherwise

RULE6 = 1, if SENG(i) is greater than ETH3 and SZCR(i) is less than VZAV+2*VZSIG

= 0, otherwise

3.2.2.2 Speech-Silence Consideration

One may say that a simple and elegant speech detector can be implemented by taking advantage of the property of the acoustic energy level difference between speech and silence. This is true if a considerable error rate is permissible, but in general, a reliable speech detector, with a correct rate above 95%, cannot be achieved by simply applying a single feature like the energy level, the bispectrum, or the zero crossing rate of the speech signal.

When real speech is used, it is very difficult (often impossible) to mark the exact point where speech starts or ends even by a careful manual inspection of a fully experienced speech scientist. Even if the speech data was collected in a noise free sound room and, as a result, had a high signal-to-noise ratio, this task is not easy. In general, it becomes more difficult to locate the beginning and the end of an utterance if there are 1) weak fricatives at the beginning or end, 2) weak plosive bursts at the beginning or end, 3) nasals at the end, 4) voiced fricatives which become devoiced at the end of words, or finally, 5) vowel sounds trailing off at the end of an utterance [2,19].

In the two-channel algorithm, problems of locating the end of an utterance with either a final nasal sound or a final vowel can be solved more easily with the aid of the EGG signal as a strong indicator of vocal

fold vibration. In other words, an interval would be classified into speech as long as there is a meaningful EGG signal. Larar [10] designed an improved two-channel endpoint detection algorithm compared to those of conventional one channel algorithm like Rabiner and Sambur's [42], which utilized the property of the zero crossing rate and the energy.

As shown in Table 3-1, the average energy level for silence ranges from 47 to 61 dB and that for speech (including voiced, unvoiced, and mixed sounds) is in the range of between 69 and 94 dB. The energy difference between speech and silence for each individual sentence is bigger than 10 dB. But we need to be careful enough not to miss the fact that SEAV+3*SESIG is less than UEAV+3*UESIG for sentence 3-a and sentence 3-e. This means that, even for clear-cut unvoiced frames, unvoiced sound energy level and silence energy level are overlapping. (Even though the amplitude distribution of speech signal is known to be similar to a gamma distribution, the assumption of a Gaussian distribution would not make any significant difference.) Another thing to notice is that the energy level for voiced sounds has a relatively large standard deviation.

The level crossing rate of the speech signal shows a very meaningful characteristic. Namely, it is always zero for the five silent frames at the beginning of each sentence. Unfortunately, this is not true when silent frames are in the middle of the sentence. Also, at either the beginning or the end of an utterance, some fricatives and vowels can have a level crossing rate of zero.

Given all the above obstacles to designing a reliable speech-silence classifier, the features of speech energy level, EGG level crossing rate, and

speech zero crossing rate are mainly used to realize the speech–silence classification. These features are usually used in combination to make the decision, rather than being used separately.

Figure 3–7 shows the ranges of features according to the frame classification for sentence 3–a. (For silence, the level crossing rate of differentiated and normalized EGG signals is not shown because it is always zero for the beginning five silent frames.) The overlapping ranges shown in this figure demonstrates that a successful speech detector with only one feature is not possible.

3.2.2.3 V–U–M Consideration

Once an input utterance is segmented into the two categories of silence or speech (whether voiced, unvoiced, or mixed sound), it is not hard to identify the unvoiced frames in the speech, because they have no vocal fold vibration. In order to detect vocal fold vibration, a threshold value of –0.5 was set on the differentiated and normalized EGG (DNEGG) signal. The threshold was used to evaluate the level crossing rate of the DNEGG signal as described in section 3.2.1. Unless the smoothed level crossing rate of the DNEGG signal has a value less than 0.6, the frame is primarily classified as an unvoiced one.

Another important feature for identifying unvoiced intervals is the zero crossing rate of the speech signal. Though it is well known that unvoiced sound has a relatively large zero crossing rate compared to that of voiced or mixed sound, setting an absolute threshold value for the zero crossing rate will not achieve an unvoiced sound identification algorithm

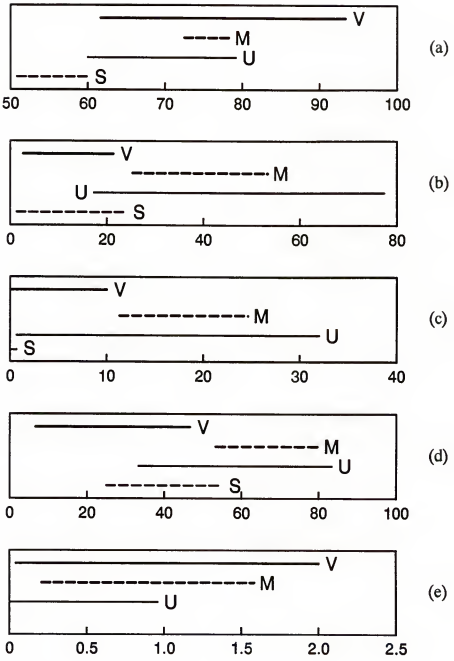


Figure 3–7. Ranges of features for Sentence 3–c: (a) speech energy, (b) zero crossing rate for speech, (c) level crossing rate for speech, (d) zero crossing rate for differentiated speech, and (e) level crossing rate for differentiated and normalized EGG

with a 100% correct recognition rate. Still, no one can deny its usefulness in detecting unvoiced intervals.

As described above, mixed sound can be considered as a combination of voiced and unvoiced sound. Hence, it is natural to assume that the five selected features usually have intermediate values compared to those for unvoiced and voiced sounds. (This is seen easily with a scan of Figure 3–7.) This makes the identification of mixed sound very difficult, and this is why a multi-level classification strategy is preferred to a single- or double-rule-based one in this study: to achieve a satisfactory mixed sound identification algorithm.

3.2.2.4 Algorithm Implementation

The underlying idea in the design of the algorithm is that different decision rules have to be applied according to the energy level of the subject sound. A somewhat less complex set of rules is necessary to classify speech segments with a high energy level. The entire range of a sentence is divided up into subranges based on the speech energy statistics, and different rules are applied to the subranges. For example, the level crossing rate of the speech signal can play an important role in identifying a voiced interval when the speech signal has a relatively large energy. But, if the same rule is applied when the speech signal has a very low energy level, a misclassification will result, because a weak voiced sound has a level crossing rate near zero.

The following is a brief description of the operating procedure of the algorithm.

Step 1: If SENG is greater than VEAV-0.5*VESIG, label the frame as "voiced".

Step 2: If SLCR is less than 0.5 and ELCR is less than 0.1, label the frame as "silence".

Step 3: If SENG is greater than SEAV+3.5*SESIG, go to Step 6.

Step 4: If SENG is less than SEAV+1.5*SESIG, label the frame as "silence".

Step 5: If RULE2 is true, i.e., equal to 1, label the frame as "unvoiced". Otherwise, label it as "silence".

Step 6: If ELCR is less than 0.6, go to Step 9.

Step 7: If RULE1 is true, label the frame as "mixed".

Step 8: If SDZCR is greater than VDZAV+2.5*VDZSIG, label the frame as "mixed". Otherwise, label it as "voiced".

Step 9: If ELCR is greater than 0.1, go to Step 12.

Step 10: If RULE3 is false, label the frame as "silence".

Step 11: If RULE4 is true, label the current frame as "unvoiced". Otherwise, label it as "voiced".

Step 12: If ELCR is less than 0.2, go to Step 16.

Step 13: If RULE5 is false, go to Step 15.

Step 14: If RULE1 is true and SENG is greater than ETH2, label the frame as "voiced". Otherwise, label it as "mixed".

Step 15: If RULE1 is true, label the frame as "voiced". Otherwise, label it as "mixed".

Step 16: If RULE1 is false, label the frame as "mixed".

Step 17: If RULE6 is true, label the current frame as "unvoiced". Otherwise, label it as "voiced".

This final algorithm has been achieved by refining a primary algorithm with the training data set of six sentences. All weighting factors appearing in the decision rules and in the decision nodes of Figure 3-6 are selected to produce a "near" optimal result with the training data set. (The word "near" is an appropriate term because optimization process in this study was heuristic, rather than mathematical. It is not unusual to use heuristic optimization in speech research, and there is no acknowledged technique of feature optimization reported when a relatively complex tree-structure pattern classification algorithm is concerned.) Specifically, each weighting factor has been changed incrementally to cover all of a predetermined range, beyond which, according to the designer's judgment, it seemed impossible to get a good result. The value yielding the best result was selected and utilized in the final classification algorithm.

3.2.3 Error Correction

The idea of error correction is almost the same as that of smoothing the result. The role of this part is to correct single frame errors and double frame errors by taking advantage of general acoustic characteristics of human speech. It is well known that an independent segment, shorter than 10 milliseconds and belonging to a different category from those of its neighboring segments, does not occur in human speech except for some unvoiced plosives. Furthermore, a voiced segment shorter than 20 milliseconds is rarely found in normal conversations. An error correction

algorithm to eliminate such classifications occasionally makes mistakes, but on the whole it contributes to improve the overall performance of the classifier. Figure 3–8 depicts the error correction algorithm used in this classifier.

The four steps in Figure 3–8 require some explanation. The single frame error correction procedure, changes the value of the current frame IVUS(i) to that of its adjacent frames, when those two frames belong to the same category, but the current frame is in a different category. For example, VUV is corrected to VVV.

The single unvoiced frame error correction step changes an unvoiced frame to a voiced one when it is between a voiced frame and a silent one. For example, SUV is corrected to SVV. This correction may increase error rate in case of very short plosives since some unvoiced plosives occasionally last only a few milliseconds, a duration shorter than one frame. Fortunately, examination confirms that the duration of unvoiced plosives in our data set is usually longer than 20 milliseconds, the length of two frames.

In the single mixed frame error correction step, a mixed frame is corrected to a voiced one under two conditions: either when the previous frame is classified as voiced and the very next two frames are unvoiced ones (VMUU to VVUU), or when the two previous frames are silent and the next is voiced (SSMV to SSVV). This correction prevents the errors which may occur during the voice-onset and -offset intervals, when transitional weak voiced sounds often look like mixed ones.

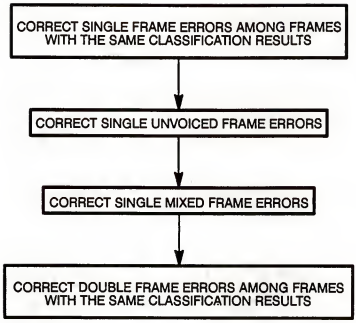


Figure 3–8. Error correction step (Two–channel)

The double frame error correction changes the values of a pair of agreeing frames to those of the immediately preceding and following pairs, when those two pairs belong to the same category but differ from that of the intervening pair. For example, SSVVSS is corrected as SSSSSS. Figure 3-9 illustrates the final classification result for the speech signal same as that shown in Figure 3-2.

3.3 Result

As shown in Table 3-2, 864 frames out of 1198 total frames in the training sentences were categorized by the "early" classification of the feature extraction step. In these results, 27 misclassified frames were found, and most of them came from either unvoiced-to-silent or silent-to-unvoiced misclassification. (Some of them also came from mixed-to-voiced misclassification because mixed sound identification was not attempted in the step.) If the unclassified frames are counted as error frames, a correct classification rate of 69.9% was obtained for the training data set after the application of the feature extraction step alone.

Table 3-3 is for the interim results obtained by applying the tree-structure pattern classification algorithm, in other words, it is the preliminary results before error correction is executed. All frames were classified and an overall 97.5% correct rate was achieved for the training data set. It may be interesting to note that a 97.2% correct rate was achieved for female speakers, while that for male speakers was 97.8%.

The final classification results for the six training sentences is depicted in Table 3-4. These results were obtained by applying the error

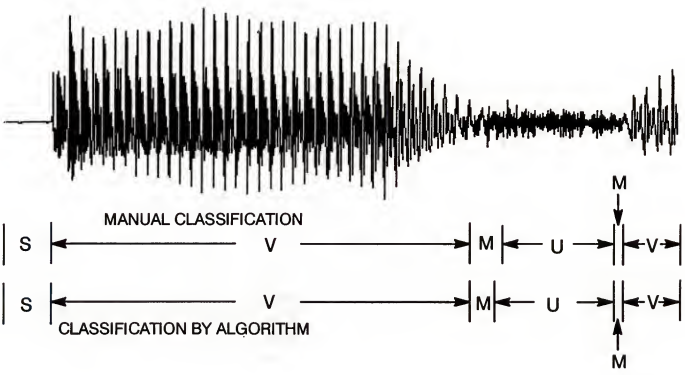


Figure 3–9. Example of four–way classification result (Two–channel)

Table 3-2. Preliminary result after feature selection (Two-channel)

SENTENCE	1-a	1-d	3-a	3-e	4-a	4-e	TOTAL
TOTAL NUMBER OF FRAMES	187	200	185	203	218	205	1198
CLASSIFIED NUMBER OF FRAMES	145	141	140	121	168	149	864
PRECLASSIFICATION RATE (%)	77.5	70.5	75.7	59.6	77.1	72.3	72.1
NUMBER OF ERRORS IN CLASSIFIED FRAMES	0	0	3	3	14	7	27
CORRECT RATE (%)	77.5	70.5	74.1	58.1	70.6	69.3	69.9

Table 3-3. Preliminary result before error correction (two-channel)

SENTENCE	1-a	1-d	3-a	3-e	4-a	4-e	TOTAL
TOTAL NUMBER OF FRAMES	187	200	185	203	218	205	1198
CLASSIFIED NUMBER OF FRAMES	187	200	185	203	218	205	1198
PRECLASSIFICATION RATE (%)	100	100	100	100	100	100	100
NUMBER OF ERRORS IN CLASSIFIED FRAMES	4	2	5	8	4	7	30
CORRECT RATE (%)	97.9	99.0	97.3	96.1	98.2	96.6	97.5

Table 3-4. Final result after error correction (Two-channel)

SENTENCE	1-a	1-d	3-a	3-e	4-a	4-e	TOTAL
TOTAL NUMBER OF FRAMES	187	200	185	203	218	205	1198
CLASSIFIED NUMBER OF FRAMES	187	200	185	203	218	205	1198
PRECLASSIFICATION RATE (%)	100	100	100	100	100	100	100
NUMBER OF ERRORS IN CLASSIFIED FRAMES	0	0	3	5	3	0	11
CORRECT RATE (%)	100	100	98.4	97.5	98.6	100	99.1

correction algorithm to the results from the classification algorithm. We can confirm the efficacy of our error correction step simply by comparing the overall correct rate of Table 3-3 to that of Table 3-4. An improvement of 1.6% in the overall correct rate was achieved during the error correction step for the training sentences.

Finally, detailed classification results appear in Table 3-5. The designations for the test data sets are as follows:

1) "COMPLETE" refers to all the data from all six speakers.
2) "THRESHOLD" is for the subset of "COMPLETE' used to establish the threshold values, one male and one female speaker, for sentence 1, 3, and 4 only.
3) "NON-THRESH." represents the subset of "COMPLETE" which is not included in the "THRESHOLD" set.
4) "MALE" refers to the male speaker subset of "COMPLETE".
5) "FEMALE" is for the female speaker subset of "COMPLETE".

The overall correct rate is 98.7%, as judged against skilled manual classification of the data. The classification rate is an improvement over the overall 95% rate reported by Rabiner et al. [6] and Siegel and Bessey [5], and the 88% rate reported by Atal and Rabiner [3]. While a nearly 83% correct classification of the mixed excitation frames was achieved by Siegel and Bessey [5], our algorithm yields a 90.1% correct rate for the identification of the mixed sounds. Male speakers produced a better classification result than females by 1.0% and the "THRESHOLD" data give a better recognition rate than the "NON-THRESH." data.

Table 3-5. Classification result (Two-channel)

TEST DATA SET	TOTAL # OF FRAMES	# OF FRAMES IN ERROR	CORRECT RATE (%)
COMPLETE	7599	99	98.7
THRESHOLD	1198	11	99.1
NON-THRESH.	6401	88	98.6
MALE	3783	29	99.2
FEMALE	3816	70	98.2

3.4 Error Analysis

In Table 3-6, the result of the error analyses in number of frames is summarized. About 25% of the misclassified frame errors occurred at the beginning and end of the sentences or at the pauses existing in the middle of the sentences. If these errors are disregarded, as is done in Siegel and Bessey's algorithm [5], then an overall performance of 99.0% would be achieved. Another type of error is that occurring in transition regions, such as voiced-to-unvoiced, unvoiced-to-voiced, voiced-to-mixed, or mixed-to- voiced intervals. These errors are considered to be caused by two main reasons. One is the problem which is inherent to our fixed frame size. If, for example, a frame consists partially of unvoiced sound and partially of voiced sound, the frame is likely to be classified as a mixed one. The other is a problem that lies in the nature of speech itself. When voiced-to-unvoiced or unvoiced-to-voiced transitions occur in human speech, the voiced sound near the boundary tends to have a very low energy level and is often being classified as unvoiced. This failure to recognize voice-onset and voice-offset intervals properly, is heavily responsible for this type of error and for similar errors regarding silent-to-voiced and voiced-to-silent transition intervals.

Excluding all above types of errors, less than 50% of the error frames are left unexplained. These includes some unvoiced-to-silent misclassification errors which seem to come from the speaker's aspiration rather than the speech itself. (It may be interesting to note that this kind of error occurs more often in female than in male speakers.) It is believed that the threshold values of our current algorithm are mainly responsible for

Table 3–6. Error analyses in number of frames (Two–channel)

MANUAL CLASSIFICATION \ CLASSIFICATION OUTPUT		V	U	M	S	CORRECT RATE(%)
V	TOTAL	5312	18	27	10	99.0
	MALE	2770	9	6	4	99.3
	FEMALE	2542	9	21	6	98.6
U	TOTAL	4	709	5	23	95.7
	MALE	1	313	3	0	98.7
	FEMALE	3	396	2	23	93.4
M	TOTAL	3	6	82	0	90.1
	MALE	2	1	53	0	94.6
	FEMALE	1	5	29	0	83.9
S	TOTAL	0	3	0	1397	99.8
	MALE	0	3	0	618	99.5
	FEMALE	0	0	0	779	100

such errors. In fact, intra-speaker variability, inter-speaker variability, and even the variability within one sentence make it impossible to set threshold values which give a 100% correct rate in general speech research.

3.5 Discussion

This chapter described a speaker-independent two-channel (speech and EGG) four-way (V/U/M/S) classification algorithm. Obviously, in many situations, the EGG signal is either unavailable or cannot be used. In the laboratory, however, both speech and EGG signals can be used to help benchmark the performance of numerous speech systems. It can easily be seen that the use of the EGG signal made our algorithm simpler and more accurate by helping the voiced-unvoiced and mixed-unvoiced classification. (Another version of a simpler two-channel four-way classification algorithm, with a 98.1 correct recognition rate, is given in Appendix.)

There is no doubt that if the EGG signal were unavailable, it would be practically impossible to achieve this overall recognition rate of 98.7%, the highest ever reported. Furthermore, a 90.1% identification rate of mixed intervals would remain as only a dream without a very powerful tool like the EGG signal. This is why we advocate the laboratory use of this algorithm (or the simpler one in Appendix), to benchmark speech system performance. The benchmarking can be done automatically and the results compared to those of algorithms based solely on the acoustic signal. A final attractive trait of our algorithm is, with minimal modifications, it can also provide endpoint information essential for the time alignment of an input word and a template in an isolated word recognition system.

CHAPTER 4

ONE-CHANNEL FOUR-WAY CLASSIFICATION

4.1 Introduction

Instead of using both the speech and the EGG signal, one-channel four-way classification algorithm utilizes only the speech signal. The EGG signal is usually unavailable in real situations and a designer has to design a speech system relied only on speech input. In this case, it is not possible to take advantage of the EGG signal as an indicator of vocal fold vibration and, as a result, the system becomes more complicated. The added complexity is mainly due to the difficulty in voiced/mixed-unvoiced/silent classification. Different features, such as spectral distribution [6,43] or the LPC (Linear Predictive Coding) error signal [2,3,5,9], have to be included in the feature set if a reliable classifier is needed. In this part of the study, the same set of 6 sentences is used as the training data set to develop a one-channel four-way classifier as was used for our two-channel classifier.

The features selected for our one-channel four-way classification algorithm are

1) speech energy,
2) zero crossing rate,
3) level crossing rate,

4) the zero crossing rate of the differentiated speech signal, and

5) spectral distribution.

As with the two-channel four-way classification, all the feature values are evaluated on frame by frame basis with a frame size of 100 data points (10 milliseconds). Only the last feature is changed from those of the two-channel classifier.

The main reason to select the spectral distribution is for its unique characteristics for each type of sound. The spectrum of voiced sounds shows that most of the energy is concentrated below 1 kHz and the first formant, usually the highest peak, is located below 350 Hz. For unvoiced sounds, most of the speech energy is found above 2.5 kHz and the highest peak is also found in this region. (Even though the first formant for unvoiced sound is usually located below 450 Hz, its energy level is lower than that of the third or the fourth formant.) In the case of mixed sounds, the spectrum is relatively flat for the whole frequency region. The examination of the spectra of mixed sounds indicates that there are usually two peaks. One is located below 1 kHz and the other above 3 kHz. It is believed that the former is produced by the low frequency carrier component (due to a vocal fold vibration) and the latter is caused by the noise-like high frequency component (due to a turbulent airflow), both of which exist in a mixed sound. In Figure 4-1, examples of spectra for voiced, unvoiced, mixed, and silent segments are shown, while a spectrogram corresponding to the speech signal in Figure 3-1 is given in Figure 4-2.

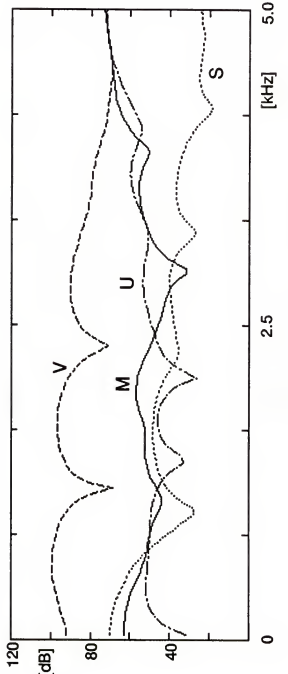


Figure 4-1. Spectral distribution of voiced, unvoiced, mixed, and silence

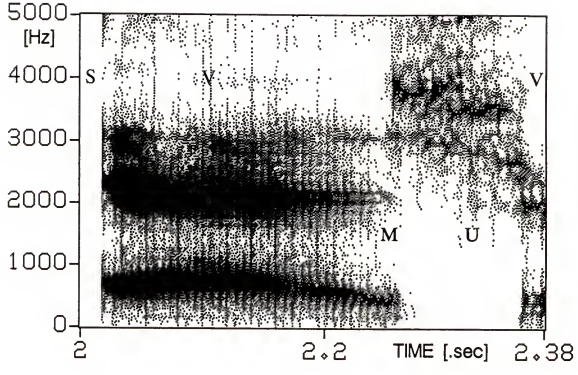


Figure 4–2. Spectrogram of the speech signal in Figure 3–2.

4.2 Algorithmic Details

The basic structure of the one-channel four-way classifier is the same as that of the two-channel four-way classifier except that it accepts only the speech signal only as its input.

4.2.1 Feature Extraction

Time-domain analysis techniques, such as zero crossing rate, energy, and level crossing rate, are not sufficient to achieve a successful one-channel four-way classification. This is clearly shown by the ranges of such features, as shown in Figure 3-6. Hence, in this study, five spectral energy ratios are added to the time-domain features to form the feature set. Details of the feature extraction step are presented in Figure 4-3.

4.2.1.1 Time-Domain Features

The time-domain features in Figure 4-3 were selected for the same reasons applicable to the two-channel four-way classifier. The statistics for each feature are evaluated in the same manner. In other words, the averages and standard deviations for these features were calculated for the "clear-cut" voiced, the "clear-cut" unvoiced, and the five beginning silent frames. The early classification of these "clear-cut" frames were achieved by applying simple rules. A frame is labeled as unvoiced, if all five spectral ratios (as defined in the next section) are less than zero. When all five ratios are greater than 20, the frame was classified as voiced. The parameter definitions were the same as those of the parameters appearing

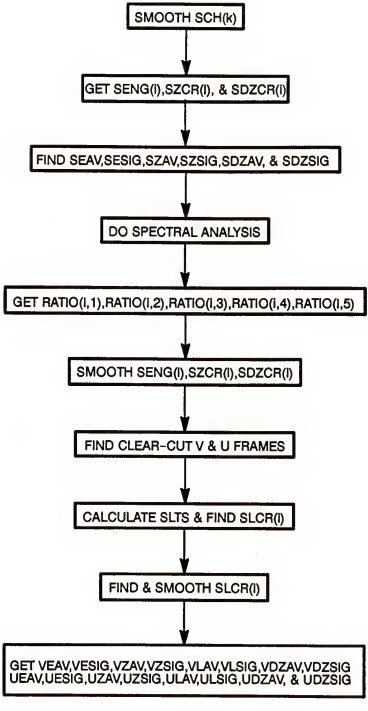


Figure 4–3. Feature extraction step (One–channel)

in the two-channel classifier, except for one threshold, SLTS, which is described as follows:

SLTS: 8% of the average magnitude of the rectified voiced sound.

As before, this threshold value was used to evaluate the level crossing rate of the speech signal.

4.2.1.2 Spectral-Domain Features

As features representing the spectral properties of the speech signal, five spectral energy ratios were selected. The selection of five, rather than a single ratio as some other researchers have done, is based on the observations that 1) even in the sentences pronounced by one speaker, a significant spectral deviation can occur even for one phoneme due to the phonetic environment, 2) the same phoneme spoken by one speaker at different times can have different spectral distributions according to the speaker's condition, mood, and intention (intra-speaker variability), and 3) for different speakers, same phoneme can have considerably different spectral distributions (inter-speaker variability). The ratios are evaluated from the spectral distributions obtained by applying the Welch method. The ratios and the threshold based on the spectral characteristics of the input speech are

RATIO(i,1): the ratio of the spectral energy of the i-th speech frame in the 150-400 Hz band to that in the 3800-4200 Hz band.

RATIO(i,2): the ratio of the spectral energy of the i-th speech frame in the 150-400 Hz band to that in the 4200-4600 Hz band.

RATIO(i,3): the ratio of the spectral energy of the i-th speech frame in the 150-400 Hz band to that in the 4600-5000 Hz band.

RATIO(i,4): the ratio of the spectral energy of the i-th speech frame in the 800-1200 Hz band to that in the 4200-4600 Hz band.

RATIO(i,5): the ratio of the spectral energy of the i-th speech frame in the 800-1200 Hz band to that in the 4600-5000 Hz band.

STH: RATIO(i,1)+RATIO(i,2)+RATIO(i,3)-RATIO(i,4)-RATIO(i,5)

It is well known that the quality of the periodogram gets worse with increasing data length. Namely, the variance of the periodogram is proportional to the data length and a frame having more than 100 points usually results in a useless periodogram for speech scientists. Hence, in order to obtain a more consistent spectrum estimate, Bartlett [44] suggested a modified version of the periodogram evaluation technique. In his method, a frame to be analyzed was divided into smaller subframes, and the spectral estimate for each segment was evaluated as the convolution of the true spectrum with the Fourier transform of the triangular window function. The final spectral estimation was obtained by averaging the periodograms for the subframes. The variance of Bartlett's estimate decreased by the factor of the number of segments, and resulted in a consistent spectral estimate.

Welch [45] has introduced a modification of the Bartlett procedure that is particularly well suited to direct computation of a power spectrum estimate using the FFT (Fast Fourier Transform). A data frame of length N is further divided into K segments having M samples. The window, $w(n)$, is applied directly to the data segments before computation of the

periodogram. The modified periodogram of the i-th segment can then be defined as

$$J_{i,M}(\omega) = 1/MU \left[\sum_{n=1}^{M-1} x_i(n) w(n) e^{-j\omega n} \right]^2 \qquad (4\text{–}1)$$

where

$$U = 1/M \sum_{n=1}^{M-1} w^2(n) \qquad (4\text{–}2)$$

and the spectrum estimate is defined as

$$B_{xx}(\omega) = 1/K \sum_{i=1}^{K} J_{i,M}(\omega) \qquad (4\text{–}3)$$

In this method, the variance of the final periodogram is also reduced by a factor of K.

In this study, the spectral distribution of the speech signal was evaluated with this Welch method and a Hamming window. The window size was 28 samples and five windows are fitted into a frame, which results in a 28.6% overlap between two adjacent windows. In order to improve the resolution of the periodogram, 228 zeros were appended before executing FFT to each windowed data set. The final resolution of our periodogram was 39.1 Hz.

As an example, the statistics for the six-sentence training data set are given in Table 4–1. In this table, the statistics for the five beginning silent frames for each sentence are not given because they are the same values as those appearing in Table 3–1. Only the statistics for the first spectral energy ratio are presented instead of listing all values for the five spectral energy ratios.

TABLE 4.1 Statistical values (One-channel)

SENTENCE / FEATURES	1-a	1-d	3-a	3-e	4-a	4-e
VEAV (VESIG) (dB)	82.9 (7.4)	90.9 (4.2)	88.4 (5.7)	90.0 (5.9)	96.3 (3.3)	89.2 (4.6)
UEAV (UESIG) (dB)	* (*)	* (*)	71.3 (3.4)	75.6 (4.2)	* (*)	* (*)
VZAV (VZSIG)	9.1 (2.8)	8.4 (2.5)	10.4 (5.0)	10.2 (4.1)	11.8 (6.4)	11.3 (4.8)
UZAV (UZSIG)	* (*)	* (*)	70.0 (10.3)	63.9 (8.6)	* (*)	* (*)
VLAV (VLSIG)	8.9 (3.1)	8.6 (2.7)	10.3 (4.7)	9.8 (3.8)	11.8 (6.5)	10.7 (4.6)
ULAV (ULSIG)	* (*)	* (*)	47.2 (15.5)	50.4 (11.4)	* (*)	* (*)
VDZAV (VDZSIG)	22.1 (7.0)	17.6 (7.1)	22.8 (9.2)	20.5 (8.6)	25.0 (7.9)	26.5 (8.4)
UDZAV (UDZSIG)	* (*)	* (*)	77.7 (6.1)	73.5 (6.9)	* (*)	* (*)
VR1AV (VR1SIG)	39.3 (4.1)	38.8 (4.4)	38.2 (5.5)	41.9 (3.3)	34.0 (7.2)	34.8 (5.4)
UR1AV (UR1SIG)	* (*)	* (*)	-10.5 (4.2)	-6.6 (3.4)	* (*)	* (*)
SR1AV (SR1SIG)	28.1 (3.2)	33.7 (4.0)	27.1 (9.7)	26.8 (2.8)	31.9 (4.0)	33.1 (2.8)

* Data unavailable due to the absence clear-cut unvoiced frames

4.2.2 Pattern Classification

In Figure 4–4, details of the tree–structure pattern classification algorithm for the one–channel four–way classifier are shown. The threshold values and decision rules in this figure were determined based on the five features (including the spectral ratios) and their statistics in the same fashion as in Chapter 3.

4.2.2.1 Threshold Explanation

If there was no clear–cut unvoiced frame in a given sentence, the threshold values were defined as

$$ETH1 = (SEAV+VEAV)/2 \quad (4–4–a)$$

$$ETH2 = (SEAV+VEAV-2*(SESIG+VESIG))/2 \quad (4–5–a)$$

$$ETH3 = (SEAV+VEAV+0.5*(SESIG+VESIG))/2 \quad (4–6–a)$$

When there were some clear–cut unvoiced frames, these values were set as follows.

$$ETH1 = UEAV \quad (4–4–b)$$

$$ETH2 = UEAV-UESIG \quad (4–5–b)$$

$$ETH3 = UEAV+UESIG \quad (4–6–b)$$

The rules in our one–channel four–way pattern classification algorithm were defined as

RULE1 = 1, if all five ratios are less than zero

= 0, otherwise

RULE2 = 1, if all five ratios are greater than 30

= 0, otherwise

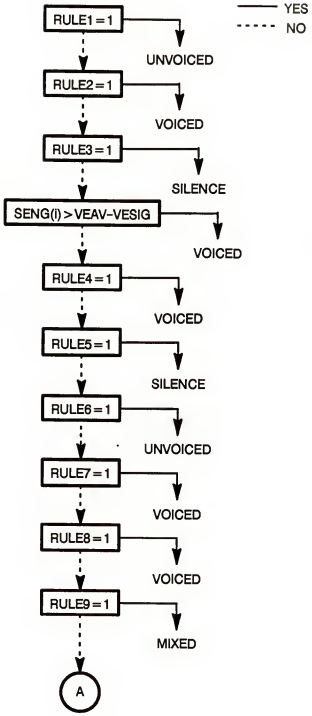


Figure 4–4. Details of pattern classification step (One–channel)

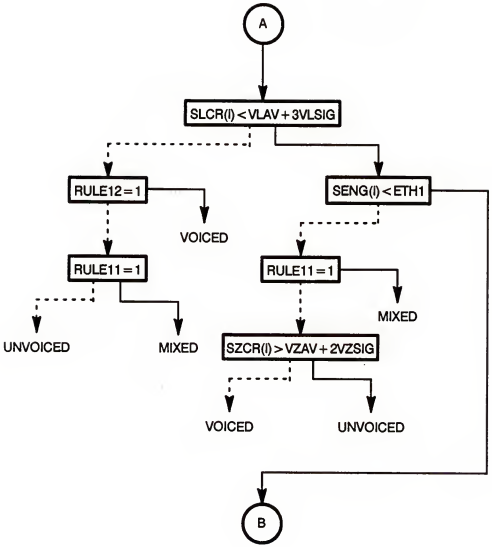


Figure 4–4.—Continued

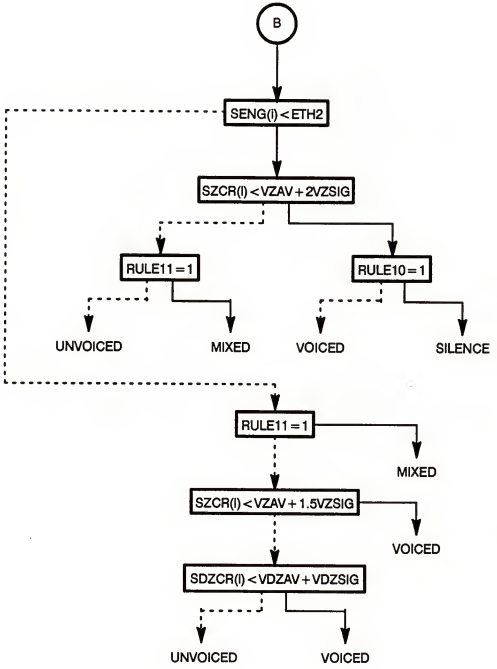


Figure 4–4.—Continued

RULE3 = 1, if SENG(i) is less than SEAV

= 0, otherwise

RULE4 = 1, if SENG(i) is greater than VEAV-2*VESIG, SZCR(i) is less than VZAV+2*VZSIG, SLCR(i) is less than VLAV+2*VLSIG, and SDZCR(i) is less than VDZAV+2*VDZSIG

= 0, otherwise

RULE5 = 1, if SENG(i) is less than SEAV+2*SESIG, SZCR(i) is greater than SZAV-2*SZSIG, SZCR(i) is less SZAV+2*SZSIG, SDZCR(i) is greater than SDZAV-2*SDZSIG, SDZCR(i) is less than SDZAV+2*SDZSIG, and RATIO(i,5) is greater than zero.

= 0, otherwise

RULE6 = 1, if SENG(i) is less than UEAV+UESIG, SENG(i) is greater than UEAV-1.5*UESIG, SZCR(i) is less than UZAV+UZSIG, SZCR(i) is greater than UZAV-1.5*UZSIG, SLCR(i) is less than ULAV+ULSIG, and SLCR(i) is greater than ULAV-1.5*ULSIG.

= 0, otherwise

RULE7 = 1, if all five ratios are greater than 20

= 0, otherwise

RULE8 = 1, if all three of RATIO(i,1), RATIO(i,2), and RATIO(i,3) are greater than 30, RATIO(i,4)+RATIO(i,5) is greater than zero, and SENG(i) is greater than ETH1

= 0, otherwise

RULE9 = 1, if the sum of RATIO(i,4) and RATIO(i,5) is less than -10, the sum of RATIO(i,2) and RATIO(i,3) is greater than 20, and SENG(i) is greater than ETH3.

= 0, otherwise

RULE10 = 1, if SENG(i) is less than SEAV+5*SESIG and SLCR(i) is less than 3

= 0, otherwise

RULE11 = 1, if RATIO(i,1) is greater than zero, RATIO(i,2) or RATIO(i,3) is greater than zero, RATIO(i,4) or RATIO(i,5) is less than zero, SDZCR(i) is greater than VDZAV+1.25*VDZSIG, STH is greater than 8, SDZCR(i) is greater than 40, and SENG(i) is greater than ETH2.

= 0, otherwise

RULE12 = 1, if SDZCR(i) is less than VDZAV+VDZSIG and SENG(i) is greater than VEAV-1.5*VESIG

= 0, otherwise

4.2.2.2 V-U-M-S Consideration

The basic idea for speech-silence classification for the one-channel classifier is the same as that of the two-channel one. Even though the spectral distribution for silence has unique properties, as shown in Figure 4-1, this information was not utilized in our study, since the characteristics of ambient noise could be different from place to place where data were collected.

In the design of the one-channel four-way pattern classification algorithm, it is necessary to depend heavily on the properties of the spectral distribution. Without referring to this information, it is almost impossible to accomplish voiced-mixed and unvoiced-mixed classifications with an acceptable correct rate. (As has already been emphasized, the time-domain features of mixed sounds always overlap those of both unvoiced and voiced sounds.) This is why the strategy for the one-channel pattern classification algorithm is different from that for the two-channel one.

For the one-channel pattern classification algorithm, a speech segment was primarily classified according to its spectral characteristics, while in the two-channel classifier the energy level of the speech signal played this major role. In both classifiers, detailed rules were applied thereafter, to make the final decision of assigning a segment to a specific category.

The basic criterion for distinguishing mixed sounds from voiced sounds is that the energy of mixed sounds is almost equally distributed over the entire frequency range, while the energy of voiced sounds is concentrated in the low frequency range. Hence, if the situation is ideal, we will get large values for all five ratios for voiced sounds while the values for mixed sounds are low, close to one. In practice, it was observed that mixed sounds were affected a great deal by their neighboring sounds, so that its spectral characteristics usually resembled those of their neighboring sounds. This phenomenon makes it extremely hard to identify mixed sounds located within a transition interval. Hence, because of the absence of the EGG signal, our one-channel pattern classification algorithm must become more complex, as depicted in Figure 4-4.

4.2.2.3 Algorithm Implementation

The pattern classification algorithm is designed to extract the voiced, unvoiced, and silent frames which are easy to classify first. The remaining frames are then classified into all four categories with detailed decision rules. The operating procedure of the one-channel pattern-classification algorithm is as follows:

Step 1: If RULE1 is true, i.e., 1, label the frame as "unvoiced".

Step 2: If RULE2 is true, label the frame as "voiced'.

Step 3: If RULE3 is true, label the frame as "silence".

Step 4: If SENG is greater than VEAV-VESIG, label the frame as "voiced".

Step 5: If RULE4 is true, label the frame as "voiced".

Step 6: If RULE5 is true, label the frame as "silence".

Step 7: If RULE6 is true, label the frame as "unvoiced".

Step 8: If RULE7 is true, label the frame as "voiced".

Step 9: If RULE8 is true, label the frame as "voiced".

Step 10: If RULE9 is true, label the frame as "mixed".

Step 11: If SLCR is greater than VLAV+3*VLSIG, go to Step 14.

Step 12: If RULE12 is true, label the frame as "voiced".

Step 13: If RULE11 is true, label the frame as "mixed". Otherwise, label it as "unvoiced".

Step 14: If SENG is less than ETH1, go to Step 17.

Step 15: If RULE11 is true, label the frame as "mixed".

Step 16: If SZCR is greater than VZAV+2*VZSIG, label the frame as "unvoiced". Otherwise, label it as "voiced".

Step 17: If SENG is greater than ETH2, go to Step 21.

Step 18: If SZCR is greater than VZAV+2*VZSIG, go to Step 20.

Step 19: If RULE10 is true, label the frame as "silence". Otherwise, label it as "voiced".

Step 20: If RULE11 is true, label the frame as "mixed". Otherwise, label it as "unvoiced".

Step 21: If RULE11 is true, label the frame as "mixed".

Step 22: If SZCR is less than VZAV+1.5*VZSIG, label the frame as "voiced".

Step 23: If SDZCR is less than VDZAV+VDZSIG, label the frame as "voiced". Otherwise, label it as "unvoiced".

As was done for the two-channel four-way classification algorithm, this final algorithm has been obtained by refining a primitive algorithm with the training data set. All weighting values were adjusted discretely, in the same manner as for the two-channel four-way classification algorithm, to produce the "near" optimal results for the training data set. The number of steps for this algorithm is increased by six from that for the two-channel one. It means about 35% more complexity was added compared to the two-channel algorithm, without guaranteeing any improvement in its final performance.

4.2.3 Error Correction

The basic principle used in the error correction step of the one-channel four-way classifier is a little different from that for the two-channel classifier even though the last two steps (single and double

frame error corrections) are the same as those for the two–channel classifier. In Figure 4–5, the procedure of the error correction step is presented.

In single voiced frame error correction, a mixed frame is corrected to voiced if one of its neighboring frames is voiced and the other is silent. This correction is based on the observation that a mixed sound which begins or ends an utterance would last more than 20 milliseconds. (Some mixed sounds only last one frame in voiced–to–unvoiced or unvoiced–to–voiced transition regions, but this case is excluded from this correction step.)

The single mixed frame error correction step tries to correct some mixed–to–voiced or mixed–to–unvoiced misclassifications. If the sum of RATIO(i,4) and RATIO(i,5) is less than zero and SDZCR is greater than VDZAV+VDZSIG, then the current voiced frame is corrected to mixed. When the sum of the RATIO(i,1), RATIO(i,2), and RATIO(i,3) is greater than 18, the sum of RATIO(i,4) and RATIO(i,5) is less than 15 but positive, and SENG is greater than ETH1, then the current unvoiced frame is reclassified as mixed. The single unvoiced frame error correction step corrects a voiced frame to unvoiced, when both RATIO(i,4) and RATIO(i,5) are less than zero.

Next is the single suspicious frame error correction step. A single voiced or unvoiced frame was corrected to the category of the following frame based on a simple distance measure, i.e., the taxicab distance measure of the zero crossing rate of the speech signal. Specifically, if |SZCR(i–1)–SZCR(i)| was greater than |SZCR(i+1)–SZCR(i)| and all three

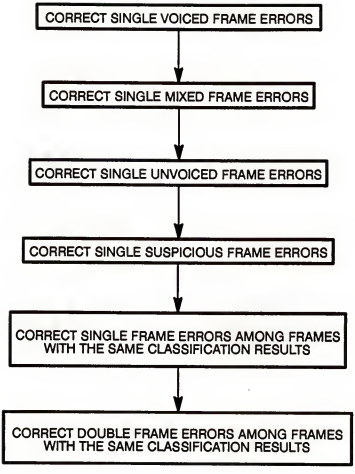


Figure 4–5. Error correction step (One–channel)

frames are classified into different categories, then the current frame is reclassified to the category of the following frame. The last two steps, single and double frame error correction are performed in the same way here as they were in the two-channel classifier. Figure 4-6 illustrates the final classification result for the same speech as that shown in Figure 3-2.

4.3 Result

Table 4-2 shows the preliminary results for the six training sentences as obtained after the feature extraction step. Of 1198 total frames, 801 frames were classified. If we count the 397 unclassified frames as error frames, a correct rate of 66.4% was obtained. This value is lower than that for the two-channel classifier by 3.5%.

In Table 4-3, the interim results obtained by applying the tree-structure pattern classification algorithm are presented. All frames were classified and an overall correct classification rate was 96.7%. This is lower than that for the two-channel classifier by 0.8%.

The final classification results for the six training sentences are given in Table 4-4. The overall performance rate was 96.9%, which is lower than that of the two-channel classifier by 1.8%. It is also found that male speech works better than female for the one-channel classifier.

In Table 4-5, we can see the results produced by applying our one-channel algorithm to the various data sets. The designations for the test data sets are the same as those described in section 3.3. The overall correct classification rate of the one-channel classifier was 96.9%. One interesting fact is that male speech yielded a 97.5% correct rate while

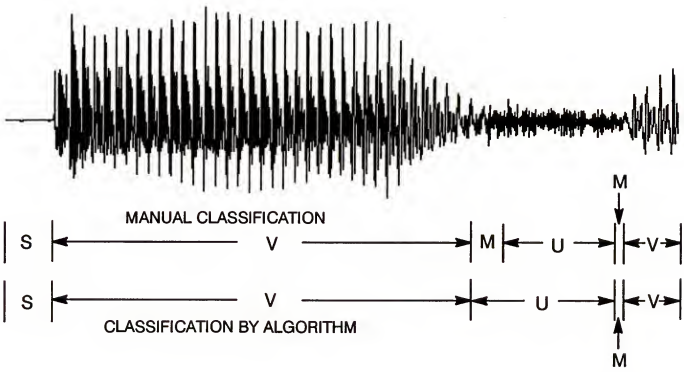


Figure 4-6. Example of four-way classification result (One-channel)

Table 4-2. Preliminary result after feature extraction (One-channel)

SENTENCE	1-a	1-d	3-a	3-e	4-a	4-e	TOTAL
TOTAL NUMBER OF FRAMES	187	200	185	203	218	205	1198
CLASSIFIED NUMBER OF FRAMES	151	152	122	102	133	141	801
PRECLASSIFICATION RATE (%)	80.7	76.0	65.9	50.2	61.0	68.8	66.9
NUMBER OF ERRORS IN CLASSIFIED FRAMES	1	0	1	3	0	1	6
CORRECT RATE (%)	79.7	76.0	65.4	48.8	61.0	68.3	66.4

Table 4–3. Preliminary result before error correction (One–channel)

SENTENCE	1–a	1–d	3–a	3–e	4–a	4–e	TOTAL
TOTAL NUMBER OF FRAMES	187	200	185	203	218	205	1198
CLASSIFIED NUMBER OF FRAMES	187	200	185	203	218	205	1198
PRECLASSIFICATION RATE (%)	100	100	100	100	100	100	100
NUMBER OF ERRORS IN CLASSIFIED FRAMES	4	2	9	13	5	6	39
CORRECT RATE (%)	97.9	99.0	95.1	93.6	97.7	97.1	96.7

Table 4–4. Final result after error correction (One–channel)

SENTENCE	1–a	1–d	3–a	3–e	4–a	4–e	TOTAL
TOTAL NUMBER OF FRAMES	187	200	185	203	218	205	1198
CLASSIFIED NUMBER OF FRAMES	187	200	185	203	218	205	1198
PRECLASSIFICATION RATE (%)	100	100	100	100	100	100	100
NUMBER OF ERRORS IN CLASSIFIED FRAMES	5	1	7	14	5	5	37
CORRECT RATE (%)	97.9	99.5	96.2	93.1	97.7	97.6	96.9

Table 4–5 Classification result (One–channel)

TEST DATA SET	TOTAL # OF FRAMES	# OF FRAMES IN ERROR	CORRECT RATE (%)
COMPLETE	7599	240	96.9
THRESHOLD	1198	37	96.9
NON–THRESH.	6401	203	96.8
MALE	3783	95	97.5
FEMALE	3816	145	96.2

female speech produced 96.2%, leaving a 1.3% gap. Thus, for both the two-channel and the one-channel classifiers, male speech gave superior results to female. The difference in the correct rate between the THRESHOLD and the NON-THRESHOLD data sets was 0.1%. From this result, we may conclude that our system adapts to the input sentences very well, and as a result is very insensitive to the change of speakers.

4.4 Error Analysis

In Table 4-6, the results of the error analyses in number of frames are summarized. About 15% of the errors were found at the beginning or ending of the sentences. At the boundaries of pauses inside the sentences, 13% of the errors were found. If we ignore these errors, caused by the failure to detect the boundaries of utterances properly, the overall correct classification rate of our one-channel four-way classifier, would be improved to 97.8%. Transition intervals, such as voiced-to- unvoiced or mixed-to-unvoiced, accounted for 38% of the total errors. The main source of this type of error could be either the fixed frame size or the improper detection of voice-onset and voice-offset, as described in section 3.4. Excluding the above two types of errors would leave 40% of the total errors unexplained. The imperfection of the features and non-optimal weighting values were believed to be mainly responsible for these errors.

In Table 4-6, a 84.1% overall correct rate is reported for unvoiced sound identification. This may be an acceptable result, but it is far from a good one. The examination of these errors showed that more than half

Table 4-6. Error analyses in number of frames (One-channel)

CLASSIFICATION OUTPUT / MANUAL CLASSIFICATION		V	U	M	S	CORRECT RATE(%)
V	TOTAL	5305	13	5	44	98.9
	MALE	2756	8	3	22	98.8
	FEMALE	2549	5	2	22	98.9
U	TOTAL	56	623	29	33	84.1
	MALE	16	279	14	8	88.0
	FEMALE	40	344	15	25	81.1
M	TOTAL	13	15	63	0	69.2
	MALE	2	11	43	0	76.8
	FEMALE	11	4	20	0	57.1
S	TOTAL	9	22	1	1368	97.7
	MALE	3	8	0	610	98.2
	FEMALE	6	14	1	758	97.3

of them were occurred either at the beginning and ending of the sentences or at the voiced-to-unvoiced transition intervals. Some of them could be removed at the expense of a poorer recognition rate for mixed sounds.

4.5 Discussion

In this chapter, a speaker-independent adaptive one-channel (speech only) four-way (V-U-M-S) classification algorithm was described. The overall performance of the system was 96.9% and a 69.2% correct rate was obtained for mixed frame identification. Comparing these two values to those of Siegel and Bessey's V-U-M classifier [5], it is found that the overall performance of our system is better than Siegel's, while Siegel's mixed frame identification rate is better than ours by 8.4%. If we recognize that Siegel obtained a 77.6% mixed frame recognition rate by testing her system only on a selected data set (16 sentences out of 48 total sentences), it could easily be said that the asserted superiority of Siegel's system in mixed frame identification will not discourage the use of our one-channel classification algorithm.

An examination of Table 3-5 and Table 4-5 shows that an 1.8% degradation in the overall system performance occurred for the one-channel classifier. (The degradation goes up to 20.9% when the mixed sound identification is concerned.) This phenomenon could be considered as an inevitable one, because in the implementation of the one-channel algorithm, we could not use the EGG signal, a powerful tool, which helped greatly in mixed sound identification and in speech-onset and speech-offset detection by eliminating some possible errors due to voice-onset and voice-offset.

This system can be used as a working criterion, either in a laboratory or in the field, to evaluate a system performance, where the EGG signal is not available. Also, it could provide endpoint information automatically with minimal modifications, which is very useful for isolated word recognition systems. For speech synthesis system, the time information needed to activate the voiced-unvoiced(-mixed) switch could be provided automatically.

CHAPTER 5
APPLICATIONS

5.1 Endpoint Detection

The problem of locating the beginning and end of a speech utterance in background noise (silence) is important in many areas of speech processing. This topic is attracting more interest recently, in conjunction with the ISDN (Integrated Services Digital Network). It is particularly essential in IWR systems to identify the speech part of the input signal, which implies endpoint (or speech) detection. Necessary computations can be significantly reduced if this identification is reliable and the extraneous data can be discarded. Furthermore, the quality of an endpoint detector will directly affect overall performance of most IWR systems, because they utilize the endpoint-based DTW (Dynamic Time Warping) technique to achieve successful time alignment between an input utterance and a stored template [10,46,47-51].

The endpoint detection problem is not trivial, except in the case of extremely high signal-to-noise ratio environments. When such high signal-to-noise ratio is guaranteed, the energy of the lowest level speech sounds exceeds the background noise energy, can thus be a reliable threshold to produce a satisfactory result. It is generally accepted that a fairly straightforward endpoint detection is possible when signal-to-noise ratio exceeds 30 dB [2]. However, such ideal conditions are not easily

met, and unfortunately, the speech data used for this study were not exceptions. They produced about 25 dB signal-to-noise ratio for voiced sounds, while that for unvoiced sounds was about 10 dB. (For mixed sounds, the ratio was about 18 dB.) From these figures, we can easily conclude that, for our speech data, it would not be possible to implement any reliable speech detector by using a simple pattern classification algorithm.

Some important examples of endpoint detection algorithms reported in the speech literature includes Rabiner and Schafer's (using speech energy and speech zero crossing rate)[2], Wilpon and Rabiner's (using Hidden Markov Model)[52], that of Lamel et al. (using energy pulses with a level equalizer) [53], and Larar's (using a two-channel algorithm) [10]. Neuberg's algorithm [54], mainly based on the low-frequency energy measurement, tried to reduce the errors occurring in voice-onset and voice-offset intervals, and could be included in this category even though it was not intended as an implementation of an endpoint detector.

In this study, the endpoint detection performance was evaluated for both the two-channel and the one-channel four-way classification algorithms. (It would be more precise to use a term, "endframe", rather than "endpoint" because our algorithm is a frame based one. The term "endpoint" is only used in order to avoid the complexity caused by using a different term from the rest of the literature.) Sixty utterances (ten digits, from "one" to "ten", uttered by three male and three female speakers) were used as a test data set.

5.1.1 Two-channel Endpoint Detection

The two-channel four-way classification algorithm was slightly modified to produce endpoint information for each input utterance. Then, the overall performance was evaluated with all sixty test words, described above. The final result is summarized and shown in Table 5-1. The designations for the types of tolerance are as follows:

1) "EXACT" counts the cases where an endpoint obtained by the algorithm precisely matched the manually determined endpoint.
2) "1-FRAME" counts the cases where an endpoint from the algorithm falls on the frame on either side of manually determined endpoint.
3) "3-FRAME" counts the cases where an endpoint from the algorithm falls within three frames on either side of the manually determined endpoint.
4) "PHONEME" counts the cases where an endpoint from the algorithm falls within the range of the beginning and end of the (speech-initial or -final) phoneme.

The selection of these types of tolerance is due to several practical reasons. For an IWR system, "EXACT" and "1-FRAME" tolerance would not cause any serious trouble when the information is used to achieve time registration between an input utterance and a stored template by an endpoint-based DTW technique. "3-FRAME" tolerance usually allows a successful registration, except when the phoneme, where the mismatch happened, has a frame length less than six. (The length of a phoneme is usually longer than six frames.) Finally, cases which fail to attain at

Table 5-1. Result of endpoint detection (Two-channel)

TOLERANCE	CORRECT RATE (%)		
	MALE	FEMALE	TOTAL
EXACT	80.0	67.7	73.3
1-FRAME	88.3	70.7	79.2
3-FRAME	93.3	81.7	87.5
PHONEME	100	90.0	95.0

least "PHONEME" tolerance, will cause a serious problem. Since the endpoint determined by the algorithm has already missed an entire phoneme of the input utterance, the time registration technique should fail to produce any useful result.

If a speech synthesis system is considered, using endpoints of "EXACT" and "1-FRAME" tolerance should not degrade the quality of output speech at all. An utterance produced with "3-FRAME" tolerance would not be hard to understand (in terms of speech science, no loss of intelligibility), but it would sound a little unnatural (loss of naturalness). When an endpoint without even "PHONEME" tolerance is used, the output speech might lose intelligibility, i.e., it might be not understandable.

Table 5-1 shows that, for use in an IWR system, our two-channel endpoint detector produced a rather successful result. In this case, the fatal errors are only those of failing to include more than half of a phoneme (three cases were counted, only in female utterances) and of missing a phoneme. Hence, if we consider our result as a general one, we can assert that for male speakers, the algorithm should be error-free in providing useful endpoint information. However, when female speakers are considered, the situation becomes significantly different. In this case, we cannot be so bold to claim perfection. We might say that the 85% is a reasonable one. Still, given a 10% error rate from missed phonemes and 5% more degradation from missing more than half of a phoneme, a disastrous errors would result within an IWR system. For a speech synthesis system, the same things can be said.

The examination of our data showed that all fatal errors described above occurred with the phonemes /f/ (in "four" and "five"), /s/ (in "six" and "seven"), /t/ (in "eight"), and /v/ (in "five"). If we recognize that almost everyone pronounces /v/ in 'five" as an unvoiced rather than a mixed sound, and that the /t/ in "eight" is usually very weak or unuttered in American English, all errors (except for one missed /s/ in "six") can be explained by the well-known difficulty of detecting weak fricatives at the beginning or end of an utterance. The unexplained error of the missed /s/ in "six" was found to be caused by the high energy level of the background noise, which was difficult to identify even by manual inspection. Another interesting thing is that all fatal errors come from two female speakers, and in both cases their speech signal shows a relatively high energy level for background noise. Overall, it is not unreasonable to expect the two-channel endpoint detection algorithm to produce valid information for every utterance, if data are collected carefully and their validity is confirmed with a somewhat more severe signal-to-noise ratio criterion.

5.1.2 One-channel Endpoint Detection

The one-channel four-way classification was also slightly modified to produce endpoint information automatically, and was tested on all sixty words. The final results are shown in Table 5-2. The same designations were used for this table as for that in the previous section. The one-channel endpoint detector was superior to the two-channel one if only the types of errors termed "1-FRAME" and "3-FRAME" were considered, but it missed one more phoneme than the two-channel detector, and if the

Table 5-2. Result of end point detection (One-channel)

TOLERANCE	CORRECT RATE (%)		
	MALE	FEMALE	TOTAL
EXACT	58.3	58.3	58.3
1-FRAME	93.3	73.3	83.3
3-FRAME	95.0	86.7	90.8
PHONEME	98.3	90.0	94.2

precise endpoint detection ("EXACT") was considered, it was much less reliable than the two-channel one.

One more interesting thing was that the same six phonemes missed by the two-channel endpoint detector were also missed by the one-channel detector. This observation supports the idea that if the data were collected carefully to yield higher signal-to-noise ratio, the endpoint detectors, both one-channel and two-channel, would produce a much better result, i.e., no fatal error for the two-channel endpoint detector and only one fatal error for the one-channel one.

5.2 Codeword Generation

As briefly described in Chapter 1, the two-pass approach in an IWR system requires that a coarse stage of recognition take place prior to the finer matching necessary for exact word identification. This step makes it possible to eliminate the unlikely words from the pool of match candidates [27,28,29]. In order to test the usefulness of the four-way classifiers, the algorithms were slightly modified to generate a codeword for each input utterances. For example, "UVSU" for an input utterance "six". Both codeword generation algorithms (two-channel and one-channel) were tested and evaluated for all sixty utterances.

5.2.1 Two-channel Codeword Generation

The output of the two-channel four-way classifier is a string of various length representing the characteristics of each 10 millisecond frame with one of four symbols. A possibility for an input utterance "six" is

SS...SSUU...UUVV...VVSSSSUU...UUSS...SS (5-1)

which includes surrounding silent frames. To describe only within word characteristics, the string can be reduced to

#U#V#S#U (5-2)

or just

UVSU (5-3)

The string in Eq. (5-2) represents acoustically homogeneous segments of duration specified by the preceding number (#). Since duration is generally rather variable, the further reduced form of Eq. (5-3) would be a more reliable representation. Hence, in this study, the representation of Eq. (5-3) is preferred.

Since we are using the data set having sixty utterances from six speakers, the maximum number of codewords for one word is six. For example, the word "six" can be represented as the codeword, UVSU, if the codeword generator is reliable and the speaker pronounced the word correctly. If the generator failed to detect the beginning /s/, then the resulting codeword would be VSU. Other conceivable variations of the codeword for "six" are UVU, VSU, UMVSU, and UMVU.

In Figure 5-1, the performance of the two-channel codeword generator is presented. The restriction to three codewords is based on practical considerations. Namely, if there are more than three variations of the codeword for one word, it can be said that the performance of the codeword generator is nearly useless. Codeword generator's performance was evaluated in the following manner. If the four codewords UVSV, UV, UVSMV, and VSV were generated for the word "seven" with the

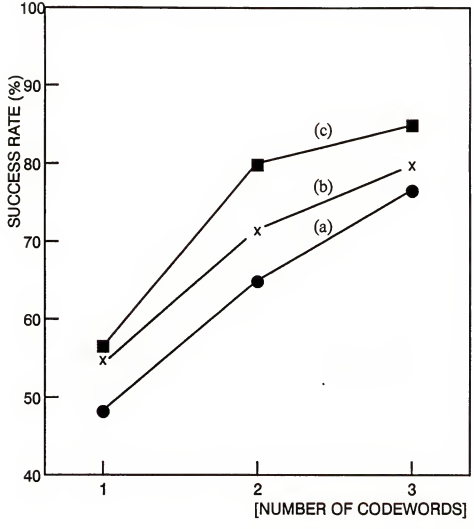


Figure 5-1. Result of codeword classification (Two-channel):
(a) codeword,
(b) codeword without mixed frame classification,
(c) codeword with two-frame-smoothing

frequency of three, one, one and one, respectively, then the test word would be recognized 50% of the time if only the most common codeword were stored in the lexicon, 67% of the time if two codewords were stored, and 83% if three codewords were stored.

There are three different results in this figure. The first result (a. codeword) was obtained with the codewords generated directly from the output of the four-way classifier without any modification, while the second (b. codeword without mixed frame classification) was obtained by changing all mixed frames to unvoiced ones. The last (c. codeword with two-frame-smoothing) was obtained by applying a two-frame-length smoother to the output string of the four-way classifier. The role of this smoother is to change a pair of frames when they are classified differently from their neighbors. For example, the string SS...SSUU...UUMMMMVV ...VSSUU...UUSS...SS would result in the codewords UMVSU, UVSU, and UMVU for conditions (a), (b), and (c), respectively.

Regardless how many codewords are permitted in the lexicon, those generated with two-frame-length smoother performed best, those without mixed frame classification were next best, and the unmodified codeword performed the worst. The lowest recognition rate was 48%, using unmodified codewords and a single lexicon entry, and the highest rate was 85%, with two-frame-length smoother and three lexicon entries.

5.2.2 One-Channel Codeword Generation

In Figure 5-2, the result of the codeword generation with the one-channel four-way classification algorithm is shown. The designations

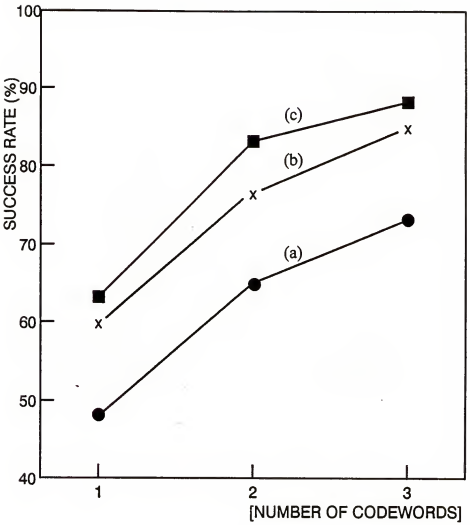


Figure 5-2. Result of codeword classification (One-channel):
(a) codeword
(b) codeword without mixed frame classification,
(c) codeword with two-frame-smoothing

and the way in which the success rates were evaluated for this figure are the same as those in Figure 5-1. The highest recognition rate of 88% was achieved using codewords with two-frame-length smoother and three lexicon entries, and the lowest, 48%, using unmodified codewords with a single lexicon entry.

The main obstacle that hindered the achievement of a high success rate was the high energy level in the background noise for both the one-channel and the two-channel codeword generators. (This was also the main obstacle for the endpoint detectors). Male utterances produced a better result than female, as expected. It was also observed that both for the one-channel and for the two-channel codeword generation, the best results were obtained from the codewords with the two-frame-length smoother, the next best came from the codewords without a mixed frame classification, and the worst from the codewords obtained directly from the output string of the four-way classifier. This observation showed two important aspects in the design of a codeword generator. They are, 1) the output of the four-way classifier had to be modified in order to produce a more reliable codeword, and 2) it would be better to use a three-way (voiced-unvoiced-silent) classifier than a four-way classifier if the concern is restricted only to the codeword generation. (It was also noted that the one-channel codeword generator was superior to the two-channel one slightly. But the difference was too small to gain any meaningful information by comparing them.)

5.3 Suggestion for Mixed Excitation

Klatt [15] suggested a pure sinusoidal signal as a glottal excitation waveform for the high quality mixed sound generation. (For mixed sound generation, both a glottal waveform and a noise source are needed.) His sinusoidal signal was shifted upward to imitate his assumption that vocal fold closure would not occur for the mixed sounds. Another technique to produce high quality mixed sound was suggested by Holm [55]. He added another noise source which provided an additional random noise signal for the vocal tract filter. He reported not only the improvement of the quality of synthesized mixed sounds but also the increase in background noise level.

Close examinations of the EGG signals in our data set used for the four-way classification showed that vocal fold closure usually still occurred in the mixed sound intervals, but the closure time was shorter than that for voiced sounds. The vocal fold opening interval was measured with the same technique as reported by Krishnarmurthy [9] and an example of that technique is shown in Figure 5-3. The measurement of the pitch period and vocal fold opening interval was done for two or three pitchs of clear-cut mixed sound, classified manually, and for three consecutive pitchs of voiced sound in a sentence. The average pitch period and vocal fold opening duration were calculated for each type of sound. Then, with these averages, the ratios of vocal fold opening durations to pitch periods for both types of sound were calculated. A test of this method on four sentences with clear-cut mixed frames showed that there was about a 25% average difference between mixed sounds and voiced sounds. For

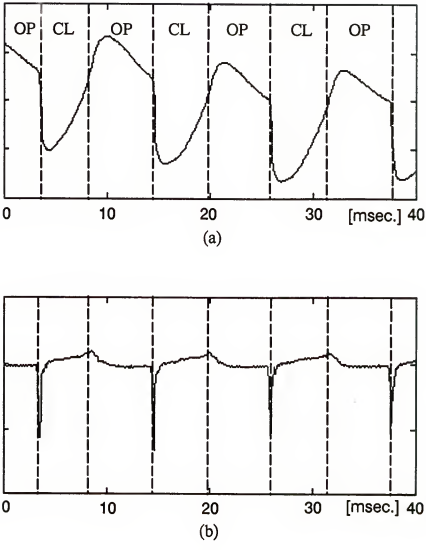


Figure 5-3. Vocal fold opening interval measurement: (a) EGG signal, (b) differentiated EGG (CL is for closing and OP is for opening)

example, if the pitch period were 10 milliseconds and vocal fold opening duration were 5 and 6 milliseconds for voiced and mixed sounds respectively, it was said that a 20% increase occurred for mixed sound. Figure 5-4 shows a typical example of this phenomenon. The T_p and T_o in this figure are for the pitch period and vocal fold opening interval, respectively.

Even though the stated value of 25% difference needs more verification, we can assert that this can be used to improve the quality of synthesized mixed sounds. In other words, the glottal excitation waveform for mixed sound has to be longer by about 25% than that for voiced sound when both sounds have the same pitch period. Another interesting observation is that in mixed sound generation, the pitch frequency usually goes down compared to the neighboring voiced sounds.

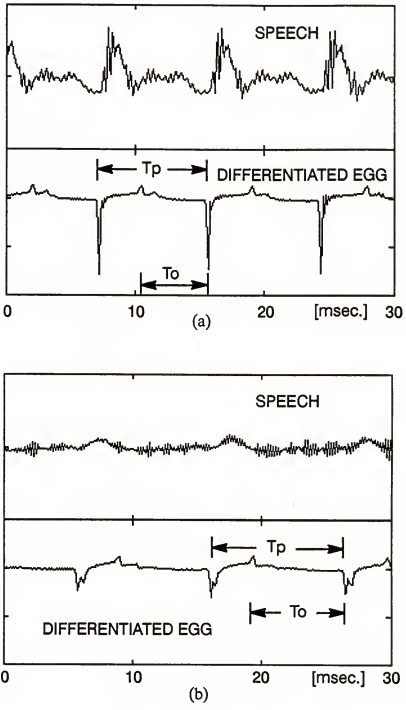


Figure 5-4. Seech and differentiated EGG signal: (a) voiced, (b) mixed

CHAPTER 6
CONCLUSION

In this study, both a two-channel (speech and EGG) and a one-channel (speech-only) four-way (voiced-unvoiced-mixed-silence) classification algorithm have been explored. A relatively simple two-channel classification algorithm was accomplished by using only time-domain features of the two signals, such as zero crossing rate and energy level. A 98.7% overall correct rate was produced in the test over all sixty sentences in the data set. For mixed sound identification, a 90.1% overall correct identification rate was achieved. These two rates are the highest correct rates ever reported. The simplicity and high quality of the two-channel classifier depended mainly on the use of the EGG signal, a strong indicator of vocal fold vibration. The one-channel classifier utilized the spectral distribution of speech to supplement the time-domain features, in order to compensate for the loss of the EGG signal. An overall correct rate of 96.9% and a mixed sound identification rate of 69.2% were achieved. Even though the performance of this classifier was worse than that of two-channel classifier, the one-channel classifier was more complex than the two-channel version because of the difficulties in identifying mixed sounds without the EGG signal.

Endpoint detectors, which are essential to most IWR systems, were designed by modifying both the one-channel and two-channel four-way

classifiers slightly to produce endpoint information automatically. The two-channel endpoint detector yielded a 95.0% overall performance when a tolerance of one phoneme was considered, while the overall performance of the one-channel endpoint detector with the same tolerance was 94.2%. It was observed that the high energy level of the background noise was mainly responsible for the endpoint detection errors.

Slight modifications of the four-way classifiers resulted in both two-channel and one-channel codeword generators, which are mainly used to reduce the number of possible word candidates in large vocabulary IWR systems. Three types of codeword were tested for all sixty words in the data set. The first was a codeword generated directly from the output string of the four-way classifier, the second was a codeword formed from the string without mixed frame classification, and the last was a codeword generated with a two-frame-length smoother. The best result was achieved with the last while the worst came from the directly generated codewords.

Finally, a suggestion was made for the glottal excitation waveform needed for mixed sound synthesis. It was based on the observation that the ratio of vocal fold opening interval to pitch period is about 25% greater for mixed sounds than for voiced sounds. Hence, for high quality mixed sound synthesis, the use of longer glottal excitation waveform was recommended.

A number of different techniques can be tested to try to improve the above systems. It may be possible to improve the four-way classifiers by assigning variable weights to different types of misclassifications. Different tree-structures may be considered for the same purpose. The effect of the

background noise level on the four-way classifiers needs to be investigated if a more practical four-way classifier is required (it was already observed that the performance of the codeword generators was severely affected by the background noise level).

There are several possible extensions of these current systems. A five-way classifier, adding a "nasal" category, could be designed with the help of the current four-way classification algorithms. This work would contribute a great deal to improving both speech synthesis and IWR systems. A new piecewise (or even linear) Dynamic Time Warping algorithm could also be developed by simply taking advantage of the boundary information which the four-way classifiers provides. This new technique would reduce the number of calculations and produce a better time alignment of the input utterance and a template. A high quality speech synthesis system could be implemented by using the voiced-unvoiced-mixed-silence information as a switching command for the excitation mode or for the parameter extraction technique. Finally, the current one-channel endpoint detector could be used in a speech interpolation system (or in the ISDN system) directly, when more refinement of the algorithm is achieved.

APPENDIX

A SIMPLE TWO-CHANNEL FOUR-WAY CLASSIFIER [56]

Another simple two-channel (speech and EGG) four-way (voiced-unvoiced-mixed-silence) classification algorithm is presented. Its algorithmic details are shown in Figure A-1. As can be seen in this figure (though the underlying idea in designing the algorithm is almost same as that for the previous two-channel algorithm), it is very simple. The algorithm was tested for the same data set of thirty sentences as was done in the previous chapters. The overall correct recognition rate for this algorithm was 98.2%, while the correct mixed frame identification rate was 89.0%. Figure A-2 shows an example of a classification result. The 0.7% degrading in the overall correct rate, compared to that of the two-channel classifier in Chapter 3, might not be a significant one in some situations. Furthermore, this algorithm is much simpler. Hence, it can also be recommended to use this classifier to benchmark speech system in laboratories.

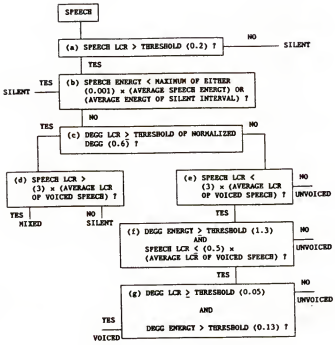


Figure A-1. A simple two-channel four-way classification algorithm

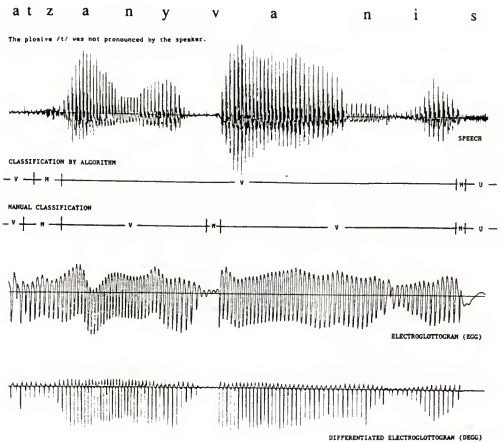


Figure A-2. Example of classification result (Simple two-channel)

LIST OF REFERENCES

1. A.K. Krishnarmurthy and D.G. Childers, "Two-channel speech analysis," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-34, no. 4, pp. 730-743, 1986.

2. L.R. Rabiner and L.W. Schafer, Digital Processing of Speech Signals, Englewood Cliffs, NJ, Prentice-Hall, 1978.

3. J.D. Markel and A.H. Gray, Jr., Linear Prediction of Speech, New York, Springer-Verlag, 1976.

4. B.S. Atal and L.R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-24, no. 3, pp. 201-212, 1976.

5. L.R. Rabiner, C.E. Schmidt, and B.S. Atal, "Evaluation of a statistical approach to voiced-unvoiced-silence analysis for telephone-quality speech," Bell Sys. Tech. J., vol. 56, no. 3, pp. 455-482, 1977.

6. F. Daaboul and J.P. Adoul, "Parametric segmentation of speech into voiced-unvoiced-silence intervals," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 327-331, New York, 1988.

7. L.J. Siegel and A.C. Bessey, "Voiced/unvoiced/mixed excitation classification of speech," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-29, no. 3, pp. 451-460, 1982.

8. L.R. Rabiner and M.R. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 323-326, Hartford, CT, 1977.

9. A.K. Krishnamurthy, Study of vocal fold vibration and the glottal sound source using synchronized speech, electroglottography, and ultra-high speed laryngeal films, Ph.D. Dissertation, Univ. of Florida, 1983.

10. J.N. Larar, Towards speaker independent isolated word recognition for large lexicons: A two-channel, two-pass approach, Ph.D. Dissertation, Univ. of Florida, 1985.

11. C.K. Un and H.H. Lee, "Voiced/unvoiced/silence discrimination of speech by delta modulation," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-27, no. 4, pp. 398-407, 1980.

12. N.J.T.M van Rossum and A.C.M. Rietveld, "A perceptual evaluation of V/U detector," Speech Communication, vol. 3, pp. 151-156, 1984.

13. B.B. Wells, "Voiced/unvoiced decision based on the bispectrum," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1589-1592, Tampa, FL, 1985.

14. P.T. Brady, "A statistical analysis of on-off patterns in 16 conversations," Bell Sys. Tech. J., vol. 47, no. 1, pp. 73-91, 1968.

15. D.H. Klatt, "Review of text-to-speech conversion for English," J. Acoust. Soc. Am., vol. 82, no. 3, pp. 737-793, 1987.

16. J.N. Holmes, "Formant synthesizers: Cascade or parallel?" Speech Communication, vol. 2, pp. 251-274, 1983.

17. J.D. Tardelli, C.M. Walter, J.T. Sims, P.A. LaFollette, and P.D. Gatewood, "Research and development for digital voice processing," Rome Air Devel. Center Tech. Rep., vol. RADC-TR-86-171, Oct. 1986.

18. D. O'Shaughnessy, "Automatic speech synthesis," IEEE Communications Magazine, vol. 21, no. 12, pp. 26-34, 1983.

19. G.J. Borden, and K.S. Harris, Speech Science Primer: Physiology, Acoustics, and Perception of Speech, Baltimore, MD, Williams & Wilkins, 1984.

20. J.L. Flanagan, Speech Analysis, Synthesis, and Perception, New York, Springer-Verlag, 1972.

21. D. Fry, Homo Loquens: Man as a Talking Animal, New York, Cambridge University Press, 1977.

22. R. Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous signals," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-36, no. 1, pp. 29-40, 1988.

23. P. D'Orta, M. Ferretti, and S. Scarci, "Phoneme classification for real time speech recognition of Italian," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 81–84, Dallas, TX, 1987.

24. P. Regel, "A module for acoustic-phonetic transcription of fluently spoken German speech," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-30, no. 3, pp. 440–450, 1982.

25. S. Makino and K. Kido, "Recognition of phonemes using time-spectrum pattern," Speech Communication, vol. 5, pp. 225–237, 1986.

26. R. Schwartz and J. Makhoul, "Where the phonemes are: Dealing with ambiguity in acoustic-phonetic recognition," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-23, no. 1, pp. 50–53, 1975.

27. J.N. Larar, "Lexical access using broad acoustic-phonetic classification," Comput. Speech Language, vol. 1, pp. 47–59, 1986.

28. D.P. Huttenlocher and V.W. Zue, "A model of lexical access from partial phonetic information," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 26.4.1–26.4.4, San Diego, CA, 1984.

29. L. Fissore, P. Laface, G. Micca, and R. Pieraccini, "Very large vocabulary isolated utterance recognition: A comparison between one pass and two pass strategy," Proc. IEEE international Conference on Acoustics, Speech, and Signal Processing, pp. 203–206, New York, 1988.

30. V.W. Zue, "Acoustic phonetic knowledge representation: Implications from spectrogram reading experiments," Edited by J.P. Haton, Automatic Speech Analysis and Recognition, Boston, MA, D. Reidel Publishing Company, pp. 101–120, 1982.

31. A.J. Fourcin, "Laryngographic assessment of phonatory function," Proceedings of the Conference on the Assessment of Vocal Pathology, Edited by C.L. Ludlow and M.O. Hart, ASHA Report, vol. 11, pp. 116–127, 1981.

32. D.G. Childers, A.M. Smith, and G.P. Moore, "Relationship between electroglottograpy, speech, and vocal contact," Folia Phoniatrica, vol. 36, pp. 105–118, 1984.

33. D.G. Childers and A.K. Krishnamurthy, "A critical review of electroglottography," CRC Critical Reviews in Bioengineering, vol. 12, pp. 131–136, 1985.

34. D.G. Childers, G.P. Moore, J.M. Naik, J.N. Larar, and A.K. Krishnamurthy, "Assessment of laryngeal function by simultaneous, synchronized measurement of speech, and ultra high speed film," Edited by L. Van Lawrence, Transcripts of the Eleventh Symposium Care of the Professional Voice, pp. 234–244, New York, 1982.

35. T. Baer, A. Lofqvist, and N.S. McGarr, "Laryngeal vibrations: A comparison between high speed filming and glottographic technique," J. Acoust. Soc. Am., vol. 73, no. 4, pp. 1304–1308, 1983.

36. Y.A. Alsaka, Electroglottographic signal analysis applied to laryngeal function assessment, Ph.D. Dissertation, Univ. of Florida, 1987.

37. K.S. Bae, Two-channel analysis: With the application to evaluation of laryngeal function and speaker identification by voice, Ph.D. Dissertation, Univ. of Florida, 1989.

38. H.K. Dunn, "The calculation of vowel resonances, and an electrical vocal tract," J. Acoust. Soc. Am., vol. 22, no. 6, pp. 740–753, 1950.

39. L.J. Siegel, "Features for the identification of mixed excitation in speech analysis," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 752–755, Washington D.C., 1979.

40. J.T. Tou and R.C. Gonzalez, Pattern Recognition Principles, Reading, MA, Addison-Wesley, 1974.

41. K.S. Fu, Synthetic Pattern Recognition and Applications, Englewood Cliffs, NJ, Prentice-Hall, 1982.

42. L.R. Rabiner and M.R. Sambur, "An algorithm for determining the endpoints of isolated utterances," Bell Sys. Tech. J., vol. 54, no. 2, pp. 297–315, 1975.

43. L.J. Siegel, "A procedure for using pattern classification techniques to obtain a voiced/unvoiced classifier," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-26, no. 1, pp. 83–89, 1979.

44. M.S. Bartlett, An Introduction to Stochastic Processes with Special Reference to Methods and Applications, New York, Cambridge University Press, 1953.

45. P.D. Welch, "The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," IEEE Trans. Audio and Electroacoust., vol. AU-15, pp. 70–73, 1967.

46. L.R. Rabiner, "Note on some factors affecting performance of Dynamic Time Warping algorithm for isolated word recognition," Bell Sys. Tech. J., vol. 61, no. 3, pp. 363–373, 1982.

47. H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP–26, no. 1, pp. 43–49, 1978.

48. L.R. Rabiner, A.E. Rosenberg, and S.E. Levinson, "Considerations in Dynamic Time Warping algorithms for discrete word recognition," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP–26, no. 6, pp. 575–582, 1978.

49. H. Sakoe, "Two–level DP–matching––A dynamic programming–based pattern matching algorithm for connected word recognition," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP–27, no. 6, pp. 588–595, 1979.

50. L.R. Rabiner and S.E. Levinson, "Isolated and connected word recognition––Theory and selected applications," IEEE Trans. Comm., vol. COM–29, no. 5, pp. 621–659, 1981.

51. S.E. Levinson, "Structural method in automatic speech recognition," Proc. IEEE, vol. 73, no. 11, pp. 1625–1650, 1985.

52. J.G. Wilpon and L.R. Rabiner, "Applications of hidden Markov models to automatic speech endpoint detection," Computer Speech and Language, vol. 2, pp. 321–341, 1987.

53. L.F. Lamel, L.R. Rabiner, A.E. Rosenberg, and J.G. Wilpon, "An improved endpoint detector for isolated word recognition," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP–29, no. 4, pp. 777–785, 1981.

54. E.P. Neuburg, "Automatic thresholding for voicing detection algorithms," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 756–758, Washington D.C., 1979.

55. S. Holm, "Automatic generation of mixed excitation in a Linear Predictive speech synthesizer," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 118–120, Atlanta, GA, 1981.

56. D.G. Childers, M. Hahn, and J.N. Larar, "Silent and voiced/unvoiced/ mixed excitation (four–way) classification of speech," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP–37, no. 11, pp. 1771–1774, 1989.

BIOGRAPHICAL SKETCH

Minsoo Hahn was born in Daejon, Korea, on November 23, 1956. He received the Bachelor and Master of Science degrees in electrical engineering from Seoul National University, Seoul, Korea, in 1979 and 1981, respectively.

After his graduation, he joined the Korea Standard Research Institute as a project engineer and carried out some projects related to pressure transducers, load cells, and automatic pressure generators.

Since 1985, he has been with the Mind-Machine Interaction Research Center at the University of Florida, Gainesville, FL. After completing the requirements for the Ph.D. degree in electrical engineering, the author is expected to join the technical staff of Yukong Software Co., in New York City.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Donald G. Childers, Chairman
Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Julius T. Tou
Graduate Research Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Jose C. Principe
Associate Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

A. Antonio Arroyo
Associate Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Nathan W. Perry, Jr.
Professor of Clinical and Health Psychology

This dissertation was submitted to the Graduate Faculty of the College of Engineering and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

December, 1989

for Winfred M. Phillips
Dean, College of Engineering

Madelyn M. Lockhart
Dean, Graduate School